UNIVERSIDAD AUTÓNOMA DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

TRABAJO DE FIN DE GRADO

Sistema de extracción de entidades y análisis de opiniones en contenidos Web generados por usuarios

Autor: Álvaro José Casado Valverde

Tutor: Iván Cantador Gutiérrez

Septiembre de 2013

Resumen

En la actualidad gran parte de los contenidos existentes en Internet son creados por usuarios finales y no por organizaciones proveedoras de contenidos y desarrolladores. Esto ha dado lugar a lo que se ha denominado Web 2.0, i.e., una Web en la que predominan sistemas como foros, redes sociales, blogs, wikis y sistemas de recomendación, todos ellos caracterizados por una alta interacción de los usuarios.

En estos sistemas los contenidos generados por usuarios poseen opiniones y sentimientos de ellos sobre lo que ocurre en el mundo real. Y es por ello por lo que surge la idea de analizar tales contenidos con el fin realizar y explotar mediciones de las opiniones y sentimientos vertidos en la Web sobre todo tipo de temas de interés. Los usos y utilidades de este análisis sobre opiniones y sentimientos extraídos de valoraciones sociales son múltiples, como por ejemplo realizar estudios acerca de tendencias de mercado, conocer valoraciones globales sobre un cierto producto o marca, o determinar la popularidad de un determinado partido político.

Con estos precedentes se plantea el presente proyecto, que tiene por finalidad el desarrollo de un sistema para la extracción de entidades y análisis de opiniones relacionadas con las mismas en contenidos generados por usuarios, y sobre una temática concreta.

En particular, como caso de uso de prueba, el sistema se ha desarrollado sobre la red social Twitter, extrayendo entidades nombradas y valoraciones de esas entidades, existentes en los mensajes cortos (“tweets”) que la gente envía en tal red social, y proporcionando resúmenes sobre tendencias de opinión (positivas y negativas) acerca de diversas temáticas de entrada.

Palabras clave

Recuperación de Información, redes sociales, extracción de entidades, minería de opinión, análisis de sentimientos.

Glosario

API *Application Programming Interface*. Librería o conjunto de funciones y procedimientos a ser utilizadas por un programa informático, consiguiendo de este modo una o varias capas de abstracción en la programación de aplicaciones finales.

Crawler Web Programa informático dedicado a la obtención de datos de páginas Web.

Follower En Twitter, usuario que "sigue" a un usuario concreto, es decir, que le establece como un usuario a cuyos mensajes quiere tener acceso y de los que quiere recibir notificaciones.

Following En Twitter, lista de usuarios a los que un usuario concreto "sigue" (del inglés *follows*).

GUI Interfaz Gráfica de Usuario (del inglés *Graphical User Interface*). Conjunto de formas y métodos que posibilitan la interacción de un sistema con los usuarios mediante el uso de componentes gráficas como iconos, botones y ventanas.

Hashtag (#) En Twitter, término precedido por el carácter '#' que se usa para resaltar o etiquetar la temática que trata el tweet generado. Ejemplo: '#educacion', referido a la temática de educación, como leyes y reformas del sistema educativo de un país.

HTML Lenguaje de Marcado Hipertextual (del inglés *HyperText Markup Language*); lenguaje de marcas con las que se escriben las páginas Web.

JSON *JavaScript Object Notation*. Formato ligero de intercambio de datos, que se caracteriza por su simplicidad de sintaxis {"objeto": valor}. Surge como alternativa a la notación XML.

Mención (@) En Twitter, término precedido por el carácter '@' que se refiere para referenciar a un usuario. Ejemplo: '@bancosantander', referido al usuario en Twitter asociado al Banco Santander.

PLN *Procesamiento del Lenguaje Natural* (del inglés *Natural Language Processing, NLP*). Disciplina de la Inteligencia Artificial que se ocupa de la formulación e investigación de mecanismos computacionales para la comunicación entre personas y máquinas mediante el uso de lenguajes naturales.

PoS *Part of Speech*. Categoría gramatical de una palabra dentro de una oración. Ejemplo: En la frase "el niño come", el etiquetado gramatical de la palabra 'el' es determinante, el de 'niño' es nombre, y el de 'come' es verbo.

SMS Servicio de Mensajes Cortos (del inglés *Short Message Service*). Servicio disponible en los teléfonos móviles que permite el envío de mensajes de texto.

Retweet (RT) En Twitter, tweet reenviado por un usuario distinto al que envío dicho tweet.

Stopword Término que recibe una palabra sin significado, como artículos, pronombres, preposiciones, etc., que son filtradas en el procesamiento de texto natural.

Tag cloud "Nube" de palabras; representación visual de las palabras que forman un texto, en la que el tamaño de fuente de una palabra da idea de la "importancia" de esa palabra dentro del texto.

ToolTip Herramienta de ayuda que se hace visible al situar el cursor del ratón sobre algún elemento gráfico.

Tweet Unidad de información de Twitter; mensaje textual publicado por un usuario en Twitter constituido por un máximo de 140 caracteres.

Twitter Sistema de *microblogging* que permite a usuarios registrados en el servicio enviar y publicar mensajes breves.

URL *Uniform Resource Locator*. Secuencia de caracteres que se utiliza para nombrar recursos en Internet para su localización o identificación. Ejemplo: http://www.uam.es, como referencia del servidor Web de la Universidad Autónoma de Madrid.

XML Lenguaje de Marcado Extensible (del inglés *eXtensible Markup Language*). Lenguaje de marcado desarrollado para el intercambio de información estructurada entre diferentes plataformas.

Índice de contenidos

Índice de figuras

1. Introducción

En este capítulo se dará una breve introducción al trabajo realizado. En el apartado 1.1 se explicarán los motivos que impulsaron la realización del proyecto. En el apartado 1.2 se detallarán los objetivos planteados en el mismo. Finalmente, en el apartado 1.3 se dará una visión general de la estructura del presente documento.

1.1 Motivación

En sus orígenes, Internet era un espacio en el que solamente unos pocos, grandes empresas en su mayoría, eran los que generaban y publicaban contenidos. Los portales Web eran páginas estáticas escritas en HTML, que se actualizaban de manera poco frecuente. Así, sus contenidos ofrecían escasa interacción con los usuarios, que eran meros "espectadores" cuya única forma de proporcionar datos era la de formularios Web. Ésta era la Web 1.0.

Con el paso del tiempo la Web fue evolucionando hasta llegar en 2004 al surgimiento de la **Web 2.0**. Este concepto debe su nombre a la lluvia de ideas que tuvo lugar entre equipos de las empresas de O'Reilly y MediaLive, durante una discusión de grupo sobre el futuro de la Web.

La Web 2.0 no sugiere una nueva "versión" de la Web desde el punto de vista arquitectónico, sino más bien una serie de cambios en la forma en que los proveedores de contenidos, desarrolladores y usuarios utilizan la Web. Estos cambios suponen que haya mayor interacción por parte de los usuarios finales, que ya no sólo consumen contenidos, sino que también tienen la oportunidad de generarlos. Es por ello por lo que la Web 2.0 se refiere a una serie de aplicaciones y servicios que utilizan los usuarios para publicar contenidos e interactuar (con esos contenidos y otros usuarios) en la Web. Así, la Web 2.0 está formada por servicios como:

- **Foros**: aplicaciones Web desarrolladas para dar soporte a discusiones o debates entre usuarios referentes a temáticas concretas.
- **Redes sociales**: servicios Web de socialización entre personas que comparten entre sí relaciones (de parentesco, personales) o gustos semejantes, o que desean explorar los intereses de otros.
- **Blogs**: también llamados bitácoras o ciberbitácoras; espacios personales en los que sus autores escriben cronológicamente artículos, información, referencias a noticias, etc. Además, otros usuarios pueden generar comentarios y opiniones acerca de las entradas publicadas.
- **Wikis**: espacios editados y mantenidos por los propios usuarios con la finalidad de compartir conocimientos. La creación y edición de contenidos se basa en la interacción de los propios usuarios a través de un navegador Web, de tal forma que múltiples autores puedan crear, modificar o eliminar contenidos identificando a cada usuario que realiza un cambio. Estas características permiten la coordinación de acciones e intercambio de información.
- **Sistemas de compartición de recursos**: sistemas Web que permiten almacenar contenidos en Internet, compartiéndolos con otros usuarios e incluso editándolos de forma colaborativa.

En estos servicios de la Web 2.0, a medida que los usuarios han ido proporcionado contenidos se ha ido produciendo una progresiva, y siempre creciente, sobrecarga de información.

Dentro de los servicios anteriores caben destacar los foros y las redes sociales, ya que en estos sistemas hay una mayor interacción social, y ya que en sus contenidos están presentes opiniones y sentimientos de usuarios sobre lo que ocurre en el mundo real. Esta información puede llegar a ser muy relevante, pues un eficaz análisis de la misma puede dar lugar a mediciones y explotación de la aceptación o rechazo en la sociedad sobre temas y cuestiones concretas. Esto es muy importante, por ejemplo en el área de los negocios para medir las fortalezas y debilidades de un determinado producto en el mercado.

Es aquí donde surge la problemática de la **minería de opinión**, mediante la cual se persigue dar una valoración cuantitativa a expresiones subjetivas asociadas a opiniones y sentimientos, y se busca identificar el grado de polaridad –positivo, negativo o neutro– en el que se califica a todo tipo de "entidades".

En este contexto, un aspecto a abordar es la propia **detección de entidades** nombradas sobre las que se vierten opiniones, ya que, además de su propia problemática de identificación dentro de un documento, conlleva la posterior clasificación temática y de opinión de dicho documento. Para esta tarea, una primera solución, que ha tenido amplia difusión, es el uso de diccionarios de entidades ya categorizadas.

La extracción de entidades y el análisis de opinión son tareas muy difíciles y costosas, todavía no resueltas, que están siendo abordadas por investigadores y empresas, y que hacen que haya una falta de herramientas comerciales de minería de opinión eficaces. A esta situación se le une el hecho de que la mayor parte de las aproximaciones existentes se han realizado para el inglés, y hay muy pocas herramientas dedicadas al español.

Debido a que no hay una solución definitiva ante este problema, en este trabajo Fin de Grado se intentará dar una aproximación al mismo, desarrollando un sistema prototipo para la extracción de entidades y análisis de opiniones sobre ellas en contenidos generados por usuarios. Este sistema además se usará como referencia en la empresa en la que el autor ha desarrollado sus Prácticas en empresa.

1.2 Objetivos

El objetivo principal de este trabajo Fin de Grado es implementar un **sistema capaz de generar resúmenes y análisis de opiniones sobre entidades asociadas a una temática dada y escritas en español**. Como punto de partida y primer caso de uso, el sistema se aplicará sobre datos generados por usuarios en una red social, concretamente Twitter[1].

Para ello, se plantean tres sub-objetivos específicos:

[1] http://twitter.com

- **Recolección de datos**. Se ha de construir un *crawler* de Twitter capaz de extraer y almacenar en una base de datos los documentos (*tweets*) de Twitter relacionados con las temáticas que se desean analizar.
- **Extracción de entidades**. Se ha de desarrollar un módulo que procese los documentos que extraídos de Twitter para extraer y categorizar las entidades nombradas en los mismos.
- **Extracción y análisis de sentimientos y opiniones**. Se ha de realizar un análisis de cada tweet con el fin de clasificar sentimientos positivos, negativos o neutros asociados a las entidades identificadas.

A través de los resultados obtenidos para los objetivos anteriores, el sistema a desarrollar será capaz de asociar y visualizar las opiniones y sentimientos globales (de los usuarios de Twitter) asociados a las entidades involucradas en una de temática de entrada dada.

1.3 Estructura del documento

La organización de este documento es la siguiente:

En el capítulo 2 se da una visión de la red social Twitter, de donde se han recolectado y analizado contenidos. Se muestran las temáticas abordadas en este trabajo, así como los términos descriptivos de cada una de las temáticas. Finalmente, se da un resumen de los datos recolectados.

En el capítulo 3 se da una visión general del sistema, y de las distintas fases de recolección y procesamiento de datos de las que consta. Además, se hace un desglose de los módulos más importantes implementados en el sistema.

En el capítulo 4 se explica la metodología llevada a cabo para la recolección del corpus de tweets, y cuestiones técnicas como el acceso a la API de Twitter para la recolección de los tweets y la planificación de descarga.

En el capítulo 5 se explica el método seguido para la extracción de entidades y su categorización, además del tratamiento previo del tweet para su normalización.

En el capítulo 6 se explican los procesos seguidos para el análisis de opinión vertido sobre una entidad.

En el capítulo 7 se explica la generación visual del análisis de opinión, representado en forma de páginas Web.

En el capítulo 8 se describen algunas de las salidas obtenidas por cada una de las temáticas tratadas por el sistema.

Finalmente, en el capítulo 9 se proporciona una reflexión y conclusiones del trabajo realizado, así como posibles mejoras que abren opciones de desarrollo de trabajos futuros.

2. Casos de uso

En este capítulo se introducirán los casos de uso abordados para probar el sistema desarrollado. En el apartado 2.1 se dará una descripción de la red social Twitter, de la que se recolectarán y analizarán contenidos generados por usuarios. En el apartado 2.2 se enunciarán las temáticas abordadas en este trabajo. En el apartado 2.3 se describirán los *hashtags* y términos identificativos de las distintas temáticas. Finalmente, en el apartado 2.4 se resumirán los datos recolectados para procesar por el sistema.

2.1 La red social Twitter

Twitter es una aplicación Web de *microblogging* muy popularizada y extendida mundialmente. Es una red social que permite que los usuarios registrados en ella escriban "tweets", mensajes de texto cortos que pueden ser leídos por cualquiera que tenga acceso a la plataforma, de acuerdo a los permisos otorgados por el usuario que escribe el tweet; un usuario puede proteger sus tweets y hacerlos únicamente visibles para ciertos usuarios concretos.

Un tweet está constituido por un máximo de 140 caracteres. Esta limitación se debe a una razón técnica histórica. Twitter fue creado antes de la popularización de los *smart phones* con acceso a Internet. En Estados Unidos los usuarios de Twitter solían actualizar sus cuentas enviando SMS a números de teléfono móvil concretos, estando el tamaño del texto de un SMS limitado a 140 caracteres. Actualmente se podrían permitir mensajes de mayor tamaño, ya que la actualización es realizada a través de Internet, pero la limitación se ha convertido en una característica diferenciadora de la red social.

A modo ilustrativo, la figura 1 muestra la página oficial en Twitter del Banco Santander[2], uno de los más importantes bancos españoles, cuya cuenta en Twitter está asociada al usuario @bancosantander:

[2] http://www.bancosantander.es

Figura 1. Página oficial en Twitter del Banco Santader, cuya cuenta es @bancosantander

En Twitter cada usuario tiene una lista de seguidores ("followers"), usuarios seguidores, que son notificados y tienen acceso a los tweets publicados por dicho usuario. Estos tweets son mostrados en la página principal del usuario. Análogamente, un usuario tiene una lista de usuarios a los que sigue ("following").

Twitter permite enviar y recibir mensajes públicos y privados entre usuarios. En los mensajes se pueden nombrar o mencionar a un usuario precediendo el nombre del usuario con el carácter '@'. Los mensajes también pueden ser clasificados mediante el uso de etiquetas que relacionen el texto con un tema determinado. Las etiquetas, denominadas hashtags, van precedidas del símbolo '#', facilitando su identificación desde el buscador.

2.2 Temáticas abordadas

Las temáticas de las que se han extraído entidades y opiniones han sido seleccionadas debido a la controversia que están generando en actualidad. Son las siguientes:

- **Reestructuración del sistema financiero en España**
  - La temática *Reestructuración del sistema financiero en España* tiene que ver con la situación actual de crisis económica, y los problemas existentes en el sistema financiero español. En este contexto, surge la necesidad de llevar a cabo en España una serie de medidas de saneamiento y reestructuración económica.
- **Bancos**
  - La temática *Bancos* es más general que la anterior. Lo que pretende es dar una visión general y actual de valoraciones hacia las principales entidades bancarias españolas.
- **Desahucios**
  - La temática *Desahucios* tiene que ver con el alzamiento o desahucio de viviendas por ejecución forzosa debido al impago de cuotas hipotecarias o de alquiler. Esta temática hace referencia a una crisis social motivada por la crisis económica.
- **Preferentes**
  - La temática *Preferentes* se refiere a la problemática de estafa por participaciones preferentes, productos financieros de alto riesgo y alta complejidad, lo que hace que sea un producto financiero muy poco recomendable para pequeños inversores a los que fue vendido.
- **Educación**
  - La temática *Educación* tiene que ver con la Educación en España, que está atravesando momentos tensos, como los recortes en becas, en profesorado, subvenciones, etc., y otros acontecimientos tales como la intención de llevar a cabo una nueva reforma educativa.
- **Sanidad**
  - La temática *Sanidad* viene marcada, al igual que las otras temáticas anteriores, por la situación actual de crisis económica, añadido a los recortes en financiación. A esto se le une la intencionalidad del Gobierno español de gestionar de forma privada la sanidad del país.

2.3 Definición de las temáticas

Para llevar a cabo la recolección de tweets hay que hacer un estudio y tratamiento previo de las temáticas a abordar. La finalidad de tal estudio es encontrar términos representativos que permitan identificar temática concretas. Basándose en los términos obtenidos, ha de hacerse un estudio de los hashtags y términos empleados en Twitter que puedan devolver resultados de búsqueda para cada temática.

Los términos empleados en la aplicación, por temática, son los siguientes:

Reestructuración del sistema financiero en España

La lista de términos de búsqueda seleccionada es la siguiente: '#sareb', '#activotoxico', '#burbujainmobiliaria', '#crisishipotecaria', '#crisisfinanciera', '#crisiseconomica',

'#bancomalo', '#banca pública', '#deuda publica', '#recapitalización', '#rescatefinanciero', '#frob'.

Desglosando por términos de búsqueda:

- El término *'#sareb'* hace referencia a las siglas Sociedad de Gestión de Activos Procedentes de la Reestructuración Bancaria[3], que es una sociedad anónima creada para gestionar activos procedentes de entidades bancarias nacionalizadas y de entidades que han requerido asistencia financiera.
- El término *'#activotoxico'* hace referencia a los activos tóxicos, que son activos financieros cuyo valor se ha reducido de manera significativa y para el que ya no es un mercado en funcionamiento, de forma que dichos activos no se pueden vender a un precio satisfactorio para el titular. Concretamente, en España tiene que ver mayoritariamente con fondos de inversión de muy baja calidad que se crean a partir de hipotecas concedidas a personas con solvencia económica baja, respaldados por una vivienda cuyo precio real difiere bastante del especulativo.
- El término *'#burbujainmobiliaria'* tiene que ver con el concepto de 'burbuja inmobiliaria' en España, que hace referencia a una tendencia que conlleva la subida de precios debido a especulaciones en el mercado de bienes inmuebles en España.
- El término *'#crisishipotecaria'* hace referencia a la crisis hipotecaria, la cual es una crisis de liquidez motivada por los impagos en cadena de deudas hipotecarias generadas en el sector de la construcción como consecuencia de la desaceleración del precio de las viviendas.
- El término *'#crisisfinanciera'* referencia a la crisis financiera que tiene como factor principal la crisis del sistema bancario, en la que activos financieros pierden gran parte de su valor nominal. Concretamente en España, las hipotecas.
- El término *'#crisiseconomica'* es equivalente al término *'#crisisfinanciera'*.
- El término *'#bancomalo'* referencia al banco malo, que es una entidad o institución financiera que se encarga de transferir los activos tóxicos de las entidades bancarias. SAREB es un banco malo.
- El término *'#banca publica'* tiene que ver con la banca pública en España, también involucrada en la crisis económica pues forma parte del sistema bancario español.
- El término *'#deuda publica'* referencia al término deuda pública, que se entiende como el conjunto de deudas que mantiene un estado frente a particulares u otro país. Se relaciona con la temática debido a que la crisis económica es el contexto de dicha situación de deuda.
- El término *'#recapitalizacion'* hace referencia a la expresión recapitalización de las entidades financieras en España, empleado para referirse al aumento de capital llevado a cabo por las entidades financieras en España, como consecuencia de las medidas de saneamiento y reestructuración del sistema financiero.

[3] http://sareb.org/

- El término *'#frob'* hace referencia a las siglas Fondo de Reestructuración Ordenada Bancaria[4], que es un fondo creado en España a partir de la crisis financiera cuyo objetivo es gestionar los procesos de reestructuración de entidades de crédito.

Bancos

La lista de términos seleccionada es la siguiente: '#bbva', '#bankia', '#bancosantander', '#bancopopular', '#kutxabanl', '#bancosabadell', '@santader_es', '@bancosantander', '@bbva', '@bankia', '@bancopopular', '@kutxabank', '@bancosabadell'.

En este caso no se hace desglose de cada uno de los términos de búsqueda, ya que cada uno corresponde a los nombres de principales bancos en España, tanto por nombre, como por nombre de usuario en Twitter.

Desahucios

La lista de términos seleccionada es la siguiente: '#hipoteca #afectados', '#hipoteca afectados', '#impago hipoteca', '#stopdesahucios', '#moroso hipoteca', '#pah', '@pah', '#dacionenpago', '#desahucio'.

Desglosando por términos de búsqueda:

- El término *'#hipoteca #afectados'* se relaciona con las personas afectadas por la hipoteca, en cuanto que no han podido cumplir los pagos de la misma, y que por tanto han sido objeto de desahucio.
- El término *'#hipoteca afectados'* es equivalente al término anterior.
- El término *'#impago hipoteca'* tiene que ver con lo que propiamente el término de búsqueda indica: el impago de la hipoteca como motivo de desahucio.
- El término *'#stopdesahucios'* hace referencia a la plataforma en contra de los desahucios StopDesahucios.
- El término *'#moroso hipoteca'* es más o menos equivalente al de '#impago hipoteca' y abarca el ámbito de la morosidad respecto a una hipoteca.
- El término *'#pah'* hace referencia a la Plataforma de Afectados por la Hipoteca[5], involucrada en los desahucios.
- El término *'@pah'* es equivalente al anterior, pero referenciando a la cuenta de Twitter.
- El término *'#dacionenpago'* tiene que ver con la dación en pago de la vivienda para afrontar la deuda contraída con una entidad financiera, en caso de desahucio.
- El término '#desahucio' hace referencia a la temática, que involucra el alzamiento o desahucio de viviendas por ejecución forzosa debido al impago de cuotas hipotecarias o de alquiler.

[4] http://www.frob.es/
[5] http://afectadosporlahipoteca.com/

Preferentes

La lista de términos seleccionada es la siguiente: 'afectados preferentes', '#preferentes'.

Desglosando por términos de búsqueda:

- El término *'afectados preferentes'* hace referencia a las personas involucradas en la estafa por las preferentes.
- El término *'#preferentes'* hace referencia al propio título de la temática, que tiene que ver con la estafa por la venta fraudulenta de acciones preferentes.

En esta temática se ha probado a realizar búsquedas de tweets con más términos, como 'acciones preferentes', 'fraude preferentes', pero se han descartado debido a que con ellos o no se obtuvieron resultados o los resultados obtenidos eran equivalentes a los obtenidos mediante la búsqueda por el término '#preferentes'.

Educación

La lista de términos seleccionada es la siguiente: '#educacion publica', '#educacion privada', '#universidad', '#wert universidad', '#wert educación'.

Desglosando por términos de búsqueda:

- El término *'#educacion publica'* hace referencia a la controversia de los recortes en educación, donde detractores de estos recortes tienen como lema: "escuela pública, de todos para todos".
- El término *'#educacion privada'* referencia a comentarios que tienen que ver con la educación privada.
- El término *'#universidad'* referencia a comentarios que tienen que ver con el ámbito universitario.
- El término *'#wert universidad'* están asociados al actual Ministro de Educación, Cultura y Deporte, José Ignacio Wert, y a los comentarios sobre él vertidos en el ámbito universitario.
- El término *'#wert educación'* al igual que el término anterior, hace referencia al actual Ministro de Educación, Cultura y Deporte, pero en este caso acotando la búsqueda a tweets que tengan el término educación.

Sanidad

La lista de términos seleccionada es la siguiente: '#sanidad', '#hospital sanidad', '#trabajadores sanidad', '#recortes sanidad', '#privado sanidad', '#publico sanidad'.

Desglosando por términos de búsqueda:

- El término *'#sanidad'* hace referencia a la propia temática, sanidad española, ya que se ha limitado la búsqueda a tweets generados en España.
- El término *'#hospital sanidad'* hace referencia a los hospitales y los posibles comentarios vertidos, acotando la búsqueda a la temática, añadiendo la palabra

'sanidad' para no obtener resultados de búsqueda que tengan que ver con los hospitales, pero no con la sanidad.

- El término *'#trabajadores sanidad'* hace referencia a los trabajadores de la sanidad, involucrados en la temática.
- El término *'#recortes sanidad'* hace referencia a los recortes presupuestarios en sanidad.
- El término *'#privado sanidad'* hace referencia a comentarios que tengan que ver con la sanidad privada.
- El término *'#publico sanidad'* hace referencia a comentarios que tengan que ver con la sanidad pública.

Como entrada del sistema ha de crearse un archivo de configuración en el cual se proveerá un diccionario de pares *clave*, como el nombre de la temática, y *valor*, como la lista con los términos que describen la temática. La figura 2 muestra un ejemplo de fichero de configuración.

```
tematicas={

'reformaFinanciera':['#sareb', '#activotoxico', '#burbujainmobiliaria', '#crisishipotecaria', '#crisisfinanciera', '#crisiseconomica', '#bancomalo', '#banca pública', '#deuda publica', '#recapitalización', '#rescatefinanciero', '#frob'],

'bancos':['#bbva', '#bankia', '#bancosantander', '#bancopopular', '#kutxabanl', '#bancosabadell', '@santader_es', '@bancosantander', '@bbva', '@bankia', '@bancopopular', '@kutxabank', '@bancosabadell'],

'desahucios':['#hipoteca #afectados', '#hipoteca afectados', '#impago hipoteca', '#stopdesahucios', '#moroso hipoteca', '#pah', '@pah', '#dacionenpago', '#desahucio'],

'preferentes':['afectados preferentes', '#preferentes'],

'educacion':['#educacion publica', '#educacion privada', '#universidad', '#wert universidad', '#wet educación' ],

'sanidad':['#sanidad', '#hospital sanidad', '#trabajadores sanidad', '#recortes sanidad', '#privado sanidad', '#publico sanidad']

}
```

Figura 2. Archivo de configuración del Crawler de Twitter desarrollado

2.4 Datos recolectados por temática

En este apartado se mostrarán algunas estadísticas de los datos obtenidos por temáticas, donde el total de tweets recolectado es de 18744. Así mismo, se darán ejemplos de tweets representativos de las temáticas abordadas, identificando algunas entidades relevantes de los mismos.

En la temática *Reestructuración del sistema financiero en España* se han obtenido un total de 311 tweets. De este conjunto de tweets, se han extraído 267 entidades, tanto menciones a usuarios (@) como entidades nombradas. En cuanto a valoraciones de sentimiento, 249 entidades aparecen valoradas como neutras, 14 como positivas y 20 como negativas entre los distintos tweets obtenidos.

Como ejemplo ilustrativo, para la entidad FROB se ha calificado su valoración como negativa en el tweet mostrado en la figura 3.

Figura 3. Tweet ilustrativo que califica la entidad FROB como o negativa

En la temática *Bancos* se han obtenido un total de 1346 tweets. De este conjunto de tweets se han extraído 1094 entidades, tanto menciones a usuarios (@) como entidades nombradas. En cuanto a valoraciones de sentimiento, 1035 entidades aparecen valoradas como neutras, 72 como positivas y 48 como negativas entre los distintos tweets obtenidos.

Como ejemplo ilustrativo, para la entidad @bancosantander se ha calificado su valoración como positiva en el tweet mostrado en la figura 4.

Figura 4. Tweet ilustrativo que califica la entidad @bancosantander como positiva

En la temática *Desahucios* se han obtenido un total de 906 tweets. De este conjunto de se han extraído 643 entidades, tanto menciones a usuarios (@) como entidades nombradas. En cuanto a valoraciones de sentimiento, 606 entidades aparecen valoradas como neutras, 29 como positivas y 54 como negativas entre los distintos tweets obtenidos.

Como ejemplo ilustrativo, para la entidad PP se ha clasificado su valoración como neutra en el tweet mostrado en la figura 5.

Figura 5. Tweet ilustrativo que califica la entidad PP como neutra

En la temática *Preferentes* se han obtenido un total de 1271 tweets. De este conjunto de tweets se han extraído 627 entidades, tanto menciones a usuarios (@) como entidades nombradas. En cuanto a valoraciones de sentimiento, 591 entidades aparecen valoradas como neutras, 19 como positivas y 30 como negativas entre los distintos tweets obtenidos.

Como ejemplo ilustrativo, para la entidad Bankia se ha clasificado su valoración como positiva en el siguiente tweet mostrado en la figura 6.

Figura 6. Tweet ilustrativo que califica la entidad Bankia como positiva

En la temática *Educación* se han obtenido un total de 10310 tweets. De este conjunto de tweets se han extraído 5191 entidades, tanto menciones a usuarios (@) como entidades nombradas. En cuanto a valoraciones de sentimiento, 4905 entidades aparecen valoradas como neutras, 397 como positivas y 219 como negativas entre los distintos tweets obtenidos.

Como ejemplo ilustrativo, para la entidad @PepeGrinan se ha calificado su valoración como negativa en el siguiente tweet mostrado en la figura 7.

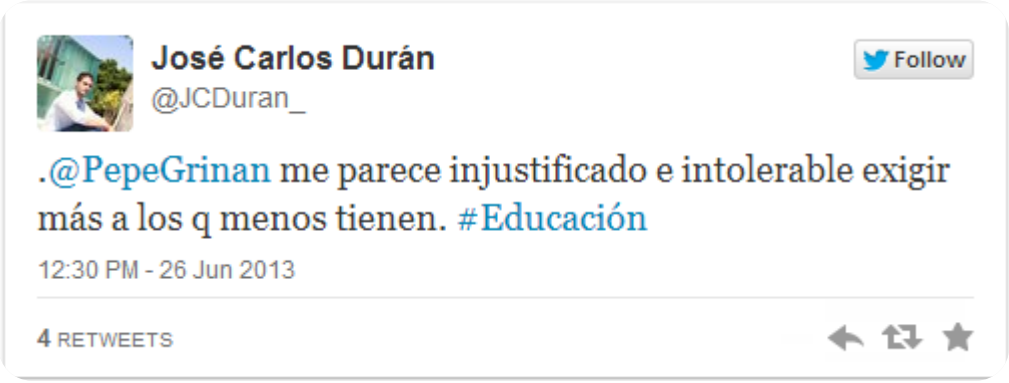


Figura 7. Tweet ilustrativo que califica la entidad @PepeGrinan como negativa

En la temática *Sanidad* se han obtenido un total de 5063 tweets. De este conjunto de tweets se han extraído 1546 entidades, tanto menciones a usuarios (@) como entidades nombradas. En cuanto a valoraciones de sentimiento, 1432 entidades aparecen valoradas como neutras, 132 como positivas y 108 como negativas entre los distintos tweets obtenidos.

Como ejemplo ilustrativo, para la entidad Lamela se ha calificado su valoración como negativa en el siguiente tweet mostrado en la figura 8.

Figura 8. Tweet ilustrativo que califica la entidad Lamela como negativa

De los datos obtenidos, se puede ver la suma del total de tweets por cada una de las temáticas es de 19207 tweets, lejos del total recolectado. Esto es debido a la existencia de tweets con coocurrencia de temáticas.

Por ejemplo, en sanidad y educación, el tweet mostrado en la figura 9:

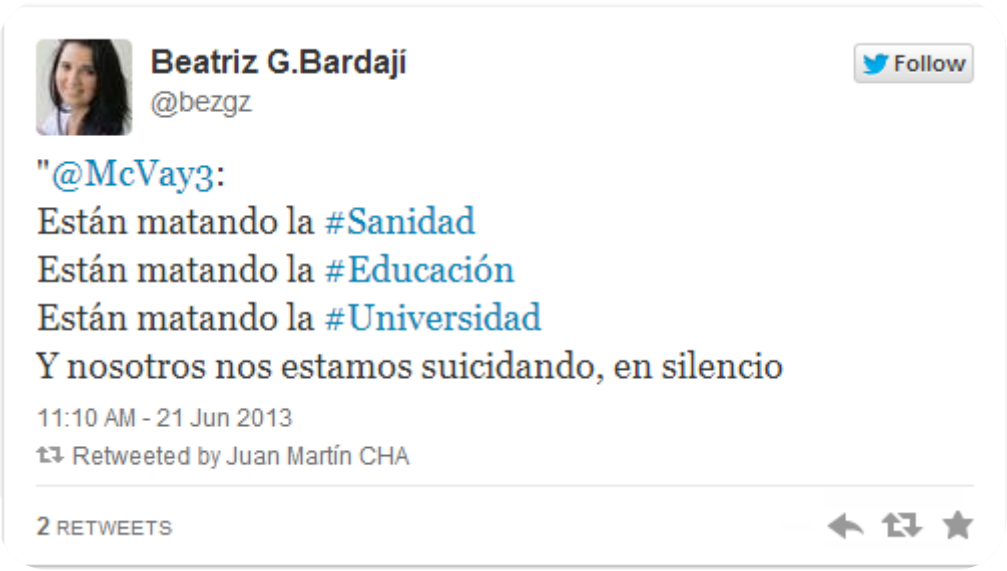


Figura 9. Tweet coocurrente para dos temáticas

3. Arquitectura del sistema

En este capítulo se dará una visión general del sistema. En el apartado 3.1 se mostrará la cadena de procesamiento seguida para el desarrollo del sistema, así como el diseño implementado del mismo. En el apartado 3.2 se dará una descripción del módulo de recolección de tweets y cómo se planifica la descarga de los mismos. En el apartado 3.3 se dará una descripción del módulo de procesamiento de texto. En el apartado 3.4 se dará una descripción del módulo de extracción de entidades y su categorización. Finalmente, en el apartado 3.5 se dará una descripción del módulo de análisis de opiniones.

3.1 Diseño arquitectónico

Para la implementación del sistema se ha optado por dividir el problema en distintas fases, de tal manera que se pueden distinguir cuatro módulos principales, cada uno de ellos dedicado a resolver una tarea específica.

La primera fase, representada en la figura 10, conlleva la **recolección de los contenidos generados por usuarios** que han de ser procesados y analizados. Para llevarla a cabo, se ha desarrollado un módulo de recolección de datos o crawler de Twitter, que se encarga de obtener el corpus de tweets mediante peticiones al servidor de Twitter, para su posterior almacenamiento en la base de datos de la aplicación.

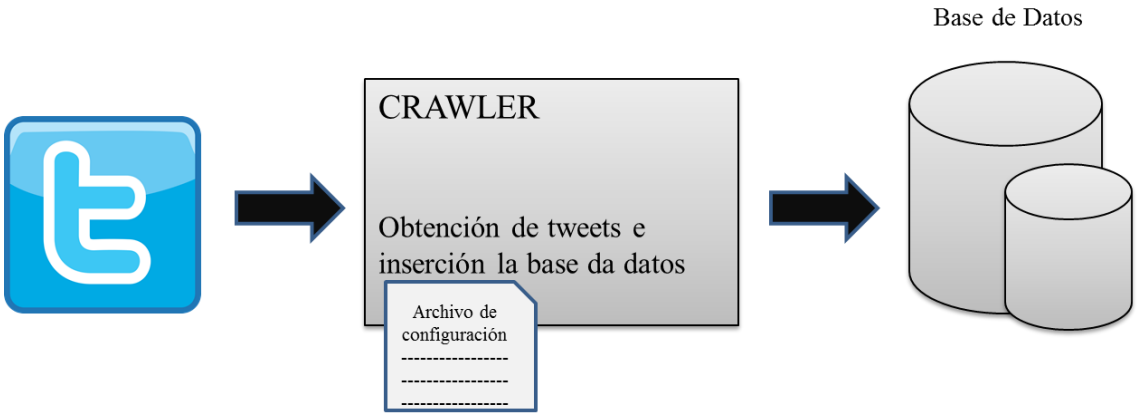


Figura 10. Arquitectura del Crawler de Twitter

La segunda fase, representada en la figura 11, es el **reconocimiento de las entidades y usuarios nombrados** en el corpus de tweets recolectados. Para llevarla a cabo, se hace necesaria una normalización previa de los textos a tratar, dado que la escritura en las redes sociales es espontánea con marcados rasgos de oralidad. En este sentido, el proceso de normalización del texto abarca tareas como corrección ortográfica y eliminación de caracteres repetidos. A partir de la normalización se puede realizar la extracción de entidades, que persigue el objetivo de obtener de cada tweet del corpus las entidades nombradas y una posterior clasificación de las mismas. Los resultados obtenidos serán almacenados en la base de datos.

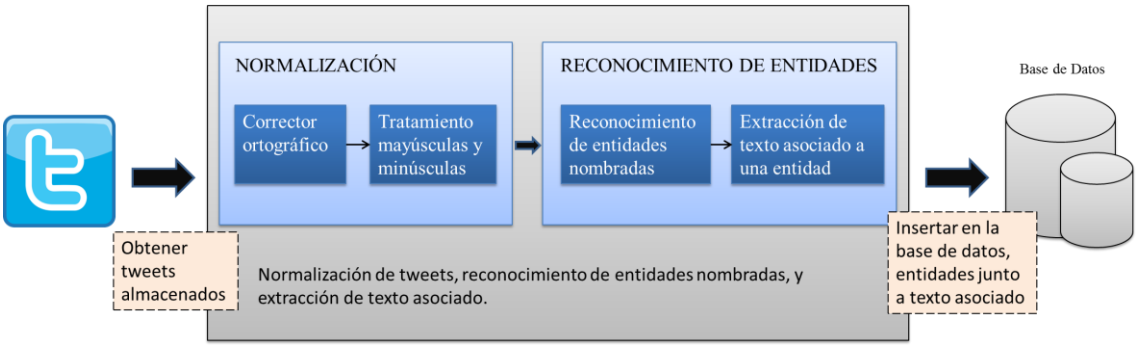


Figura 11. Arquitectura Reconocimiento de Entidades Nombradas

La tercera fase, representada en la figura 12, es el **análisis de opinión de las entidades reconocidas**. En este caso, no hace falta realizar una normalización del texto, pero si llevar a cabo un proceso de lematización por cada uno de los términos que aparecen relacionados con la entidad o usuario nombrado. Donde se extrae una valoración clasificada como positiva, negativa o neutra, según los resultados del análisis.

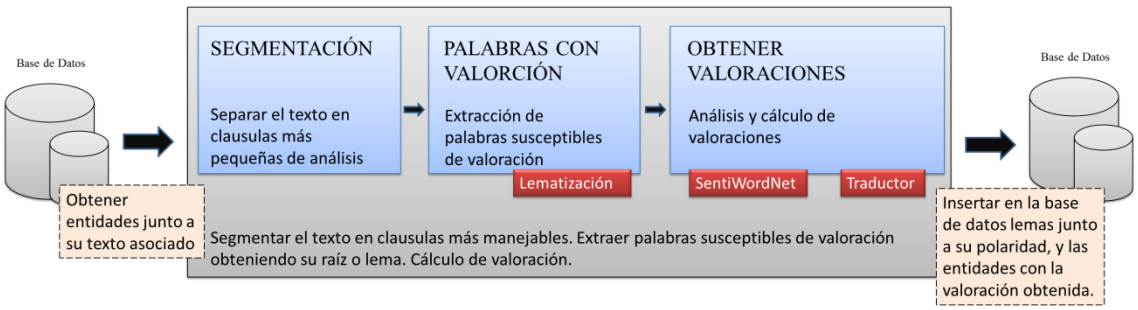


Figura 12. Arquitectura Análisis de opiniones

La cuarta y última fase, representada en la figura 13, es la de **visualización de resultados obtenidos en el análisis**. Para llevarla a cabo, se genera una Web estática con las valoraciones por entidades y por temática, con enlaces a los tweets clasificados según la valoración de la propia entidad presente.

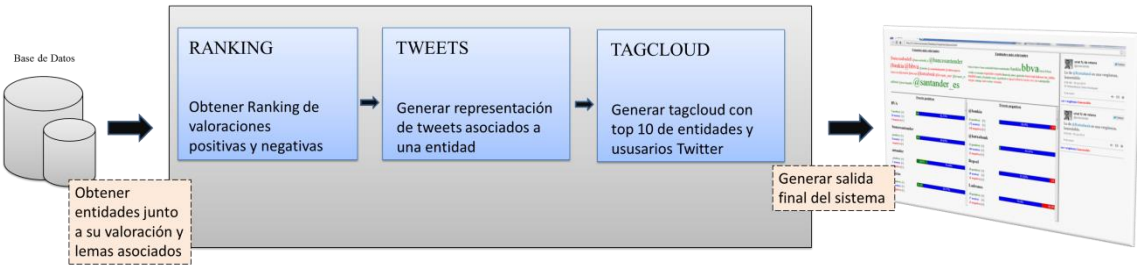


Figura 13. Arquitectura Generación de análisis de opinión

Los módulos de Reconocimiento de entidades y Análisis de opinión se nutren de un módulo externo de PLN y PoS tagging. En este contexto, cabe destacar el módulo de normalización, que realiza tareas de corrección ortográfica de los tweets, previamente al reconocimiento de entidades.

3.2 Módulo de recolección de datos

El módulo de recolección de datos consiste en un programa desarrollado en Python[6] que toma como entrada un archivo de configuración con los temas a tratar y una descripción de los mismos mediante términos que determinen la temática.

Haciendo uso de la API proporcionada por Twitter, realiza una descarga planificada de tweets, nutriendo la base de datos de la aplicación. Antes de la inserción en la base de datos, se ha de realizar una unificación de caracteres de manera que el texto del tweet almacenado tenga codificación UTF-8.

La planificación de descarga se realiza mediante crontab[7], un administrador de procesos en segundo plano proporcionado por el sistema operativo Unix.

[6] http://www.python.org/
[7] http://crontab.org/

3.3 Módulo de procesamiento de texto

El módulo de procesamiento de texto tiene como finalidad eliminar ruidos del texto. Para ello realiza una corrección ortográfica de los términos mal escritos, y lleva a cabo una eliminación de caracteres repetidos. Para las correcciones ortográficas la aplicación hace uso de la librería Aspell[8], proporcionada por GNU, mediante la cual se detectan las palabras mal escritas, y se propone una lista de palabras de sustitución.

Otro apartado a tener en cuenta es la escritura en mayúsculas, ya que interfiere en nuestra detección de entidades. Por ello, se pasa el texto a minúsculas, teniendo en cuenta los nombres, que se dejarán con la primera letra en mayúscula. Esto no afectará mucho a los resultados finales si un nombre no es reconocido como entidad nombrada, ya que aparecerá en muchos documentos y por tanto no se tomará como relevante.

3.4 Módulo de extracción de entidades

El módulo de extracción de entidades se apoya y hace uso del módulo de procesamiento de texto para el análisis de un texto limpio y su pre-procesado. Una vez limpio, se utiliza la API de Freeling[9], que cuenta con un módulo de reconocimiento de entidades dividido a su vez en varias secciones. La primera sección detecta las entidades basándose en la detección de secuencias de palabras en mayúscula, mediante lo que se consigue tener en cuenta nombres de entidades, como por ejemplo "Banco de España". Y una segunda sección trata las entidades en función de un diccionario de entidades.

No sólo se toman como entidades las detectadas por Freeling, sino que además se detectan los usuarios de Twitter mencionados en un tweet. Por cada entidad y usuario detectado se extrae y almacena el texto asociado, para posteriormente obtener su valoración.

Una vez detectadas las entidades éstas se clasifican. Para el proceso de categorización se ha tenido en cuenta el proyecto llevado a cabo por Wikipedia[10], de categorización de sus artículos. En base a esto, la categorización de entidades se realizará mediante las categorías asignadas a una entidad en Wikipedia.

3.5 Módulo de análisis de opiniones

El módulo de análisis de opiniones se centra en los textos asociados a cada una de las entidades o usuarios de Twitter mencionados. Para estos textos se hace una segmentación (*chunking*) en fragmentos más pequeños de análisis a partir de signos de puntuación y conjunciones. A fin de tratar las negaciones (e.g. "No es bueno") que invertiría la polaridad de la cláusula, se toma la valoración de cada una de las palabras de forma inversa. Fijándose en el ejemplo, "bueno" tendría polaridad positiva, pero el "no" que le precede invierte su polaridad, y por tanto "bueno" se tomaría como polaridad negativa.

[8] http://aspell.net/
[9] http://nlp.lsi.upc.edu/freeling/
[10] http://www.wikipedia.org/

4. Recolección de datos

En este capítulo se explicará la metodología llevada a cabo para la recolección del corpus de tweets. En el apartado 4.1 se dará una explicación del funcionamiento de acceso a la API de Twitter, cuyo propósito es conseguir realizar la descarga. En el apartado 4.2 se explicarán los métodos a seguir para la realización de peticiones de tweets al servidor de Twitter. Finalmente, en el apartado 4.3 se describirá cómo se realiza la planificación de descarga de tweets.

4.1 Acceso a la API de Twitter

La obtención de tweets se realiza a través de la API de Twitter[11], que da opción a realizar búsquedas de tweets mediante diferentes criterios de búsqueda.

Para acceder a la API de Twitter a través de la aplicación creada es necesario registrar y autenticar la aplicación con los servidores de Twitter. Por ello, el primer paso para conseguir la descarga de tweets es registrar la aplicación, mediante lo cual se obtendrán la *consumer key* y el *consumer secret*, que permiten la autenticación de la aplicación, así como el *access token* y el *access token secret*, que permiten la autenticación del desarrollador de la aplicación. Luego, es necesario autenticarse tanto a nivel de aplicación, como a nivel de usuario. Dicha autenticación se realiza a través del protocolo OAuth[12], estándar abierto que define un mecanismo para que una aplicación (cliente) pueda acceder a la información de un usuario en otra (proveedor) sin tener que informar a la primera del nombre de usuario y contraseña, en este caso, Consumer y Access. Dicho de otro modo, OAuth es una metodología para identificación mediante APIs genérica.

En un principio, al desarrollo de la aplicación, se estudió este modo de acceso a la API de Twitter, pero se encontraron otras formas más fáciles de acceso, como la API en Python que permite el acceso de una forma bastante sencilla. El nombre de la API es python-twitter, y el acceso se simplifica a una solo línea de código:

```
api = twitter.Api(consumer_key=CONSUMER_KEY, consumer_secret=CONSUMER_SECRET,
access_token_key=ACCESS_KEY, access_token_secret=ACCESS_SECRET)
```

4.2 Búsqueda de tweets

Una vez se ha autenticado, se procede a la búsqueda de tweets de acuerdo al archivo de configuración mencionado anteriormente, que contiene los criterios de búsqueda por términos según la temática. La búsqueda se realiza a partir de la autenticación previa, mediante la API python de acceso a la API de Twitter, siendo la búsqueda también tan sencilla como la autenticación, en una sola línea de código, mediante la llamada al método GetSearch del objeto api.

[11] https://dev.twitter.com/
[12] http://oauth.net/

```
tweets = api.GetSearch(term=term, lang= ‘es’, geocode= ‘40.426181,-3.685139,700km’)
```

Con ella se obtiene una respuesta en formato JSON por parte del servidor, pero gracias a la API desarrollada en Python, lo que se obtiene es una lista de tweets, habiendo tratado ya la respuesta JSON, con campos como el identificador del tweet, usuario que ha creado el tweet, lenguaje, texto, etc. Por cada tweet obtenido, se trata el texto y se unifica su codificación a UTF-8, y tras esto, es almacenado en la base de datos del sistema.

Respecto a la búsqueda, además de por términos, se limita la búsqueda por geolocalización únicamente a España, lo que conlleva indicar unas coordenadas de localización, por longitud y latitud, centradas en Madrid y un radio de 700Km. Para más restricción se limita el lenguaje al español.

La obtención de tweets tiene limitaciones temporales, ya que la API de Twitter no tiene indexados todos los tweets para su búsqueda, lo que conllevaría un índice de un tamaño intratable debido a la gran cantidad de contenidos generados por los usuarios. Por esta razón, se limita el índice de búsqueda de tweets a entre 6 y 9 días. La limitación no es solo temporal; también hay limitaciones de frecuencia entre peticiones de búsqueda, ya que si no podría conllevar una denegación de servicio por parte del servidor de Twitter debido a una posible sobrecarga de peticiones. Por tanto el número máximo de peticiones de búsqueda que se pueden realizar es de 450 por cada 15 minutos.

4.3 Planificación de la descarga de tweets

Dado que hay limitación en cuanto a la frecuencia de peticiones de tweets, la descarga de los mismos ha de ser programada con una cierta periodicidad. Esto conlleva a ejecutar el programa de descarga cada cierto tiempo. Como parte del diseño se ha decidido que sea cada dos horas para dejar un margen de generación de tweets.

La planificación de ejecución de la descarga se realiza mediante *crontab*, un administrador de tareas en segundo plano. Los procesos que han de ejecutarse en segundo plano, en este caso la descarga, se han de especificar en un archivo de configuración crontab, indicando los minutos, horas, días del mes, mes, día de la semana, y el comando a ser ejecutado. Los valores que sean indiferentes, por ejemplo porque se vaya a ejecutar todos los días, se deja el carácter ‘*’.

En el caso de la aplicación de descarga, el archivo de configuración crontab contendrá la siguiente línea:

```
00 0,2,4,6,8,10,12,14,16,18,20,22 * * /home/alvaro/apps/twitterApp/downloadTweets.py
```

Con esto se consigue que el programa de descarga se ejecute cada 2 horas.

5. Extracción de entidades

En este capítulo se explicará en detalle el método seguido para la extracción de entidades. En el apartado 5.1 se explicarán los criterios seguidos para el procesado del texto. En el apartado 5.2 se explicará la forma seguida en la categorización y etiquetado gramatical. Finalmente, en el apartado 5.3 se describirá el mecanismo de identificación y categorización de las entidades.

5.1 Procesamiento de texto

El lenguaje escrito utilizado en redes sociales presenta un gran desafío para el procesamiento del lenguaje, al tratarse de una escritura espontánea, con rasgos de oralidad, en su mayoría coloquial. Esto implica la necesidad de corregir errores ortográficos, como faltas de acentuación, letras repetidas, ausencia de alguna letra, etc.

En el sistema de análisis desarrollado se han tenido en cuenta estos factores, y por tanto se ha desarrollado un módulo dedicado a la corrección de errores ortográficos. Los criterios seguidos se resumen en los siguientes sub-secciones.

5.1.1 Eliminación de caracteres repetidos

En textos informales, como son los textos a analizar en la aplicación, la intensidad emotiva de las palabras implica una repetición de caracteres y especialmente en el caso de las vocales (e.g. "holaaaa, qué tal"). Estas palabras no pueden ser reconocidas, computacionalmente hablando. Por ello se ha implementado un control basado en expresiones regulares que reemplace la aparición de dos o más caracteres repetidos por uno solo, exceptuando cifras y los grupos cc, ll, rr.

5.1.2 Corrección ortográfica

Dado que la escritura de tweets es en la mayor parte de los casos espontánea y coloquial, implica errores de escritura, palabras incompletas, etc., que hacen que un ordenador no sea capaz de reconocerlas. Para minimizar este impacto se ha desarrollado un corrector ortográfico.

El corrector implementado se apoya en la herramienta GNU Aspell, la cual ha de encontrarse instalada en el equipo de procesado. Aspell no es una librería Java a la que se pueda acceder fácilmente mediante lenguaje tal lenguaje, lo que hace necesario desarrollar una interfaz de acceso a la aplicación Aspell a través de Java. La interfaz desarrollada se basa en Java Native Interface[13] (JNI), que es un framework de programación que permite que un programa desarrollado en java pueda ejecutar funciones C y C++.

El funcionamiento de Aspell es el siguiente: dado un texto, la aplicación comprueba la ortografía de las palabras basándose en el idioma de entrada. Por cada palabra ortográficamente incorrecta, se propone una lista de palabras similares para corrección. Se puede ver una propuesta de palabras de corrección en la figura 14, donde por ejemplo, para la palabra mal

[13] http://docs.oracle.com/javase/6/docs/technotes/guides/jni/

escrita “eto”, se propone como lista de palabras posibles para corrección: esto, ero, ato, feto, jeto, meto, etc.

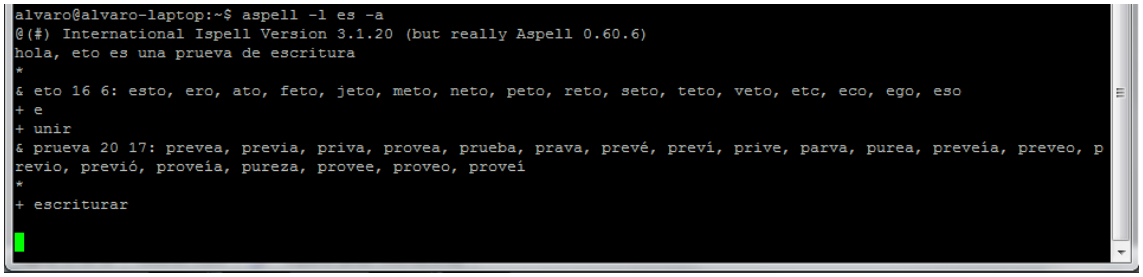


Figura 14. Propuesta corrección de palabras Aspell

De la lista de palabras proporcionada, la sustitución se basa en la mínima distancia del algoritmo Levenshtein. Este algoritmo calcula la distancia entre dos palabras como el camino mínimo de cambios (inserción, eliminación o sustitución de elementos) que han de hacerse en una cadena de caracteres para obtener la otra con la que comparamos.

A continuación se muestra un ejemplo de la corrección de un texto:

> Texto: “Wert no dbe saber que la universdd pública suspende como medda recaudatoria #educación marca #España”
>
> Corrección: “Wert no debe saber que la universidad pública suspende como medida recaudatoria #Educación marca #España”

5.1.3 Procesamiento de mayúsculas y minúsculas

Otra problemática que surge en contenidos generados por usuarios es la escritura en mayúsculas que podría significar hablar en voz alta, intentando dar énfasis a los contenidos.

Mediante la escritura en mayúsculas se pierde información, ya que todo el texto se encuentra al mismo nivel. Con ello se complica la tarea de extracción de entidades, que se basa en el reconocimiento sucesivo de palabras capitalizadas en su primera letra.

El primer paso es reconocer un tweet escrito en mayúsculas. El criterio seguido es que al menos el 60% de las palabras del texto estén en mayúsculas, omitiendo así posibles menciones o hashtags en minúsculas.

Una vez reconocido un texto en mayúsculas, éste ha de pasarse a minúsculas teniendo en cuenta los nombres presentes en el texto, ya que pueden representar una entidad. El paso a minúsculas hace necesario un análisis del texto, obteniendo la categoría gramatical de cada una de las palabras. Se hace un filtrado por cada palabra, pasándola a minúscula, a excepción de los

nombres que se dejará la primera letra en mayúscula y el resto en minúsculas. Con ello se consigue una solución aproximada al problema de la escritura en mayúsculas.

Texto: "EL DÍA 21 YA NO HABRÁ VUELTA ATRÁS. SE CANJEARÁN #preferentes POR ACCIONES Y HABREMOS PERDIDO TODO. PIDAMOS TODOS LA PARALIZACIÓN".

Minúsculas: "el día_21 ya no habrá Vuelta atrás. se canjearán #Preferentes por Acciones y habremos perdido todo. pidamos todos la Paralización."

5.2 Extracción de categorías gramaticales

La extracción de categorías gramaticales es una parte del Procesamiento del Lenguaje Natural (PLN) que consiste en conseguir un etiquetado gramatical de cada una de las palabras presentes en un texto, basándose tanto en su definición como en su contexto, es decir, en relación con las palabras adyacentes y cercanas en una frase.

1) El cogió el mando para cambiar de canal
2) Yo no mando nada.

La misma palabra "mando" se escribe y se dice igual en ambas frases, pero no significa lo mismo, ya que pertenece a contextos distintos. En la primera frase, la categoría gramatical a la que pertenece es a la del sustantivo, mientras que en la segunda sería verbo, del infinitivo mandar.

Una aproximación de etiquetado sería buscar una palabra en diccionario y categorizar en función de su definición, pero, como se ha visto, puede haber ambigüedades y el etiquetado puede ser erróneo. Por ello, dada la complejidad del problema, se ha decidido utilizar un módulo ya entrenado que tenga en cuenta la contextualización de las palabras.

El módulo externo empleado es Freeling[14], cuyo etiquetado gramatical se basa en el etiquetado gramatical HMM (Hiden Markov Model). El etiquetador HMM utiliza un modelo oculto de Markov para encontrar la secuencia de etiqueta más probable para cada oración.

[14] http://nlp.lsi.upc.edu/freeling/

El etiquetado POS usando un modelo de Markov oculto puede ser considerado como un ejemplo de inferencia bayesiana, en el que se observa una secuencia de palabras y la necesidad de asignarles la secuencia más probable de etiquetas gramaticales.

En la figura 15 se puede ver un ejemplo de etiquetado gramatical, PoS, proporcionado a través de la interfaz Web del proyecto Freeling.

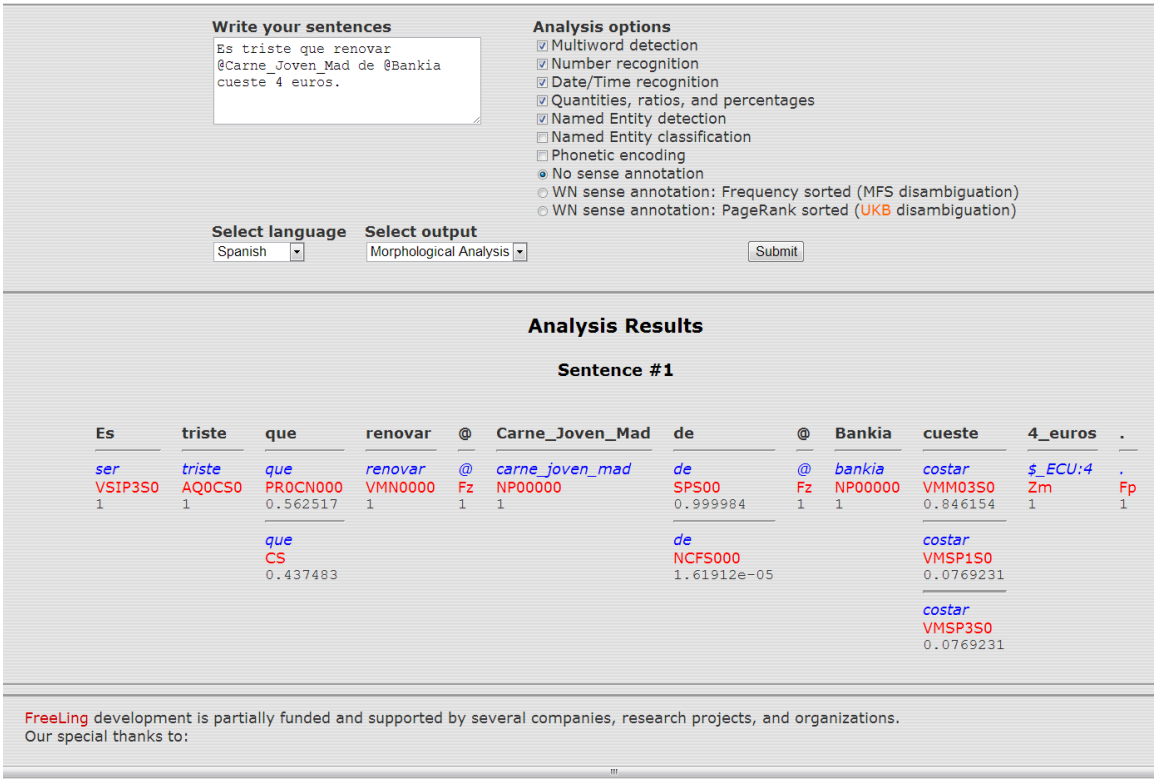


Figura 15. Etiquetado PoS, imágen de la demo de Freeling

5.3 Identificación y categorización de entidades

El Reconocimiento de Entidades Nombradas (del inglés *Named Entity Recognition*, NER) tiene como objetivo identificar y clasificar las entidades presentes en un texto, tales como nombres propios, compañías, instituciones, etc. El NER es fundamental para muchos sistemas PLN, especialmente de extracción de información.

En el sistema desarrollado, el reconocimiento de entidades es un apartado sumamente importante, ya que es uno de los principales objetivos a conseguir: obtener las entidades involucradas en una temática.

La extracción de entidades en Twitter es una tarea compleja dado al sentido coloquial presente en los textos, como las faltas ortográficas. Pero para esto se ha dado como solución el módulo de procesamiento presentado anteriormente. Una vez procesado el texto, se procede al análisis y extracción de entidades, que ha sido delegada a un módulo externo ya mencionado antes: Freeling. Éste, aparte de realizar las tareas de POS, también realiza tareas de NER.

El módulo NER de Freeling es una máquina de estados finita que básicamente detecta secuencias de palabras en mayúsculas, teniendo en cuenta algunas palabras funcionales (e.g. Banco de España) y el uso de mayúsculas al comienzo de una oración.

El archivo que controla el comportamiento del NER consta de las siguientes secciones:

- **Sección Palabras Funcionales**: lista una serie de palabras que pueden estar presentes en un nombre propio, como pueden ser preposiciones y artículos (e.g. San Sebastián de los Reyes).
- **Sección Puntuaciones Especiales**: lista etiquetas POS de signos de puntuación después de los cuales una palabra que comience en mayúsculas puede indicar el comienzo de una frase, y no necesariamente una entidad nombrada. Los casos típicos son: dos puntos, apertura de paréntesis, punto y guion.
- **Sección Etiquetado de entidad nombrada**. Contiene una única línea con el etiquetado POS que será asignado a las entidades reconocidas: NP00000.
- **Sección Ignorar**. Contiene una lista de palabras y etiquetas gramaticales que no han de ser considerados como una entidad nombrada aunque aparezcan en mayúsculas en medio de una frase.
- **Sección Nombres**. Contiene una lista de lemas que pueden ser nombres, incluso si entran en conflicto con algunos de los criterios heurísticos utilizados por el reconocedor **de entidades**.
- **Sección Afijos**. Contiene una lista de palabras que pueden ser parte de una entidad, tanto como prefijo, como sufijo. Por ejemplo, la palabra "don" en "don Juan", se consideraría como entidad "don_Juan".
- **Sección Expresiones Regulares**. Las expresiones regulares NombreAdjetivo, Cerrado y FechaNúmeroPuntuación permiten modificar las expresiones regulares utilizadas en el proceso POS. Estas expresiones regulares son utilizadas en el NER para determinar cuándo una palabra al comienzo de una frase es nombre o adjetivo, si una etiqueta es una categoría cerrada, o una fecha, número o puntuación.

Por lo que se puede ver en lo anterior, la detección de entidades es un proceso complejo en el que se aplican bastantes reglas y en el que se tienen en cuenta bastantes apartados. Esto conlleva una importante tarea de desarrollo, investigación e inversión de tiempo. Es por ello que se haya decidido el uso de este reconocedor de entidades ya consolidado, y no implementar uno propio, que tendría bastantes carencias.

Una vez se ha conseguido el reconocimiento de una entidad, comienza el proceso de categorización. Para ello se ha tenido en cuenta el proyecto "Categorías" de Wikipedia, cuyo propósito es que todos los artículos de Wikipedia han de tener al menos una categoría. En base a esto, la categorización de una entidad se ha de obtener en función de la categorización de Wikipedia. El acceso a un artículo en Wikipedia es sencillo; basta con añadir a la URL de Wikipedia: "http://es.wikipedia.org/wiki/" el nombre de la entidad a buscar. Por ejemplo, si la entidad es "FROB", el acceso al artículo se hará a través de la URL "http://es.wikipedia.org/wiki/FROB", siendo las categorías de FROB las siguientes:

Categorías: Sistema financiero Español | Banca en España | Crisis económica de 2008 en España

La obtención de dicha clasificación se consigue gracias a la librería JSOUP[15], que provee los recursos necesarios para realizar las peticiones al servidor de Wikipedia obteniendo así el artículo correspondiente a la entidad. También provee un *parser* HTML, con el que se consigue obtener las categorías. Entre peticiones se provoca que la aplicación ejecute un *sleep* de un segundo, para que el servidor de Wikipedia no deniegue el acceso a los contenidos.

En el caso de que una entidad precise desambiguación, se realiza una búsqueda en Wikipedia a través de la URL http://es.wikipedia.org/w/index.php?search=termino&title=Especial%3ABuscar, sustituyendo "término" por el nombre de la entidad a buscar. De todos los resultados obtenidos y ordenados por relevancia, se accede al primero que en su enlace contenga el nombre de la entidad. Por ejemplo, en la búsqueda de "Guindos", se accederá al artículo "Luis de Guindos".

5.4 Extracción del texto asociado a una entidad

La identificación de una entidad o un usuario Twitter no ha de centrarse únicamente en esa tarea; además ha de extraer el texto asociado para su posterior valoración. Ya que en un mismo tweet puede haber varias frases, y no todas pueden estar hablando de una única entidad o usuario.

La solución propuesta consiste en segmentar el texto en cláusulas, segmentando el tweet por signos de puntuación, paréntesis, interrogaciones, exclamaciones. Por conjunciones coordinantes, como y, e, o, o bien, u, aunque, en cambio, más bien, no obstante, pero, o sea. Y por conjunciones subordinantes como a donde, donde, a fuerza de, en vista de que, mientras, según, como quiera que, visto que, ya que, así, así que, porque, por, si bien.

De cada clausula se reconocen las entidades y los usuarios Twitter, y a su vez se va almacenando en una variable auxiliar el texto asociado, hasta que en una siguiente cláusula aparezca otra entidad o usuario.

Por ejemplo, para el siguiente texto de un tweet:

[15] http://jsoup.org/

"Hola Lamela y Güemes, estáis de mierda hasta el cuello, y ¿quiere el @ppmadrid seguir privatizando nuestra #Sanidad? http://www.eldiario.es/_8ad76a5"

se obtendría para cada una de las entidades o usuarios reconocidos, la siguiente lista de textos asociados:

Lamela: [Hola Lamela y Güemes, estáis de mierda hasta el cuello]

Güemes: [Hola Lamela y Güemes, estáis de mierda hasta el cuello]

@ppmadrid: [quiere el @ppmadrid seguir privatizando nuestra #Sanidad]

Una vez identificada una entidad o usuario, y su texto asociado, se almacena en la base de datos de la aplicación para su posterior análisis.

6. Análisis de opiniones

En este capítulo se explicarán los procesos seguidos en el análisis de opinión sobre una entidad o usuario Twitter. En el apartado 6.1 se explicará el tratamiento previo del texto asociado a una entidad o usuario. Finalmente, en el apartado 6.2 se explicará cómo se extrae la valoración asociada a una entidad o usuario.

6.1 Extracción de palabras susceptibles de valoración

La extracción de palabras susceptibles de valoración consiste en un proceso previo a la valoración de una entidad.

Por cada una de las entidades o usuarios Twitter almacenados en la base de datos junto a su texto segmentado en cláusulas, se trata su texto eliminando en una primera pasada las *stopwords*. Las stopwords son una lista de palabras carentes de sentido por sí solas, como: a, ahí, al, allí, ambos, ante, aquel, etc. Con lo que se consigue eliminar ruido en el proceso de análisis.

El siguiente paso es la "lematización", que consiste en extraer la raíz o lema de cada una de las palabras. La lematización es llevada a cabo mediante Freeling, que a través del análisis de texto, extrae el lema y el etiquetado gramatical POS de cada una de las palabras. En este proceso se han de tener en cuenta las negaciones no y ni, ya que implican un cambio en la polaridad de sentimiento. Por ejemplo, "me cae bien" implicaría sentimiento positivo, dado que bien implica positivo, pero en cambio, "no me cae bien" invertiría la polaridad de sentimiento a negativo. La segmentación facilita este proceso, ya que el análisis se hace a partir de cláusulas de la oración y no la oración entera, por lo que una negación no afecta a la inversión de sentimiento de toda la

frase, sino solamente al de la cláusula. Toda esta información se almacena en una lista de la clase Lema, la cual contiene el lema de la palabra, la etiqueta POS y un campo booleano inversión, que indica si la polaridad ha de invertirse.

6.2 Obtención de valoraciones

Una vez se han conseguido las palabras susceptibles de sentimiento, se han de procesar y medir la valoración de carga subjetiva. La cual se consigue gracias a SentiWordNet. SentiWordNet es un recurso léxico ideado expresamente para dar apoyo a la clasificación de sentimientos en minería de opinión. SentiWordNet asigna a cada *synset* (grupo de palabras sinónimas) de WordNet tres puntuaciones de sentimiento: positivo, negativo y neutro. La presentación de los términos en SentiWordNet consiste en un fichero plano, indicando el POS de una palabra y las puntuaciones negativas y positivas.

POS	ID	PosScore	NegScore	SynsetTerms	Gloss
a	00001740	0.125	0	able#1	(usually followed by `to') having the necessary means or skill or know-how or authority to do something; "able to swim"; "she was able to program her computer"; "we were at last able to buy a car"; "able to get a grant for the project"
a	00002098	0	0.75	unable#1	(usually followed by `to') not having the necessary means or skill or know-how; "unable to get to town without a car"; "unable to obtain funds"
n	00001740	0	0	entity#1	that which is perceived or known or inferred to have its own distinct existence (living or nonliving)
n	00002137	0	0	abstraction#6 abstract_entity#1	a general concept formed by extracting common features from specific examples

Para contextualizar, WordNet es una base de datos con términos en inglés, donde nombres, verbos, adjetivos y adverbios se encuentran agrupados en conjuntos de sinónimos (synsets). WordNet es una herramienta útil para el PLN, para procesos tales como POS. El único problema que se plantea es que estos recursos se encuentran solamente en inglés.

El módulo externo empleado para el análisis y etiquetado gramatical, Freeling, se nutre del recurso EuroWordNet, que se basa en WordNet para dar soporte a más idiomas, como el español y el catalán. Aun así, el recurso SentiWordNet solamente provee términos en inglés, que por tanto se han de traducir.

Para obtener el score de un lema dado en SentiWordNet, se ha de leer el fichero y almacenar cada uno de los términos, categoría gramatical en un diccionario, donde el score será una media de las puntuaciones totales. Ahora que tenemos los términos, categoría gramatical en memoria, surge el problema del idioma. Para acceder a un lema antes se ha de traducir del español a inglés.

La traducción se realiza mediante un diccionario online inglés-español, y al igual que con las categorías de Wikipedia, el acceso a los resultados se hace mediante JSOUP, obteniendo una lista de posibles traducciones. Las traducciones son almacenadas en un HashMap al que se accede antes de realizar alguna petición al servidor del diccionario inglés-español.

Solventado el problema de idioma, el acceso a un término se basa en pasar al diccionario SentiWordNet el primer término de la traducción, y si no se encuentra, el siguiente, y así sucesivamente hasta obtener el score almacenado. En caso de que el término no se encuentre, se devuelve cero. Ante la lista de palabras se plantea un umbral de neutralidad, que ante un análisis de un amplio conjunto de palabras, se ha establecido entre -0,33 y 0,33 siendo tomadas como neutras las palabras cuyo score se encuentra en ese intervalo, estableciendo un score de cero.

Habiéndose conseguido estos pasos intermedios, la valoración de una entidad o usuario consistirá en la media de todos los lemas asociados, estableciendo un umbral de opinión neutra. En este caso, se ha establecido entre -0,06 y 0,06. Por debajo de -0,06 se considerará negativo, por encima de 0,06 positivo, y entre ambos valores neutro.

Ejemplo para el texto: “Es triste que renovar @Carne_Joven_Mad de @Bankia cueste 4 Euros mientras que en otras CCAA es gratis. Luego @ComunidadMadrid presumiendo...”. En la figura 16 se puede ver las entidades obtenidas, junto a su texto asociado y el score de valoración obtenido.

```
Entidad -> @carne_joven_mad
Es triste que renovar @Carne_Joven_Mad de @Bankia cueste 4 euros.
-0.08321

Entidad -> @bankia
Es triste que renovar @Carne_Joven_Mad de @Bankia cueste 4 euros.
-0.08321

Entidad -> CCAA
en otras CCAA es gratis. Luego @ComunidadMadrid presumiendo....
0.0

Entidad -> @comunidadmadrid
Luego @ComunidadMadrid presumiendo....
0.0
```

Figura 16. Ejemplo de análisis de valoración para una entidad

Por cada entidad y texto asociado, se saca una valoración negativa para @Carne_Joven_Mad, @bankia. Y neutra para CCAA y @comunidadmadrid. La polaridad de cada uno de los lemas que han generado las distintas valoraciones son las siguientes:

[Es triste que renovar @Carne_Joven_Mad de @Bankia cueste 4 euros]

LEMA: ser NEUTRO 0.0

LEMA: triste NEGATIVO -0.41605839416058393

LEMA: renovar NEUTRO 0.0

LEMA: costar NEUTRO 0.0

[en otras CCAA es gratis, Luego @ComunidadMadrid presumiendo...]

LEMA: ser NEUTRO 0.0

LEMA: gratis NEUTRO 0.0

LEMA: luego NEUTRO 0.0

LEMA: presumir NEUTRO 0.0

Como se puede ver en el texto anterior, los lemas obtenidos califican “triste” como negativo, ya que su score está por debajo del umbral -0.33, y para el resto de lemas del ejemplo, se puede ver que han sido calificados como neutros. Se les ha asignado un score de 0, dado que su score está comprendido en el rango [-0.33, 0.33], establecido como neutro.

7. Generación de análisis de opinión

En este capítulo se dará una explicación del resumen del tratamiento de los tweets. En el apartado 7.1 se mostrarán los criterios seguidos para generar un ranking de valoraciones. En el apartado 7.2 se explicará cómo se representan los tweets presentes en la Web. En el apartado 7.3 se detallará la creación de un *tag cloud* de entidades y usuarios Twitter. Finalmente, en el apartado 7.4 se dará una visión general de la salida Web del sistema.

7.1 Representación de valoraciones

Una vez obtenidas todas las entidades y usuarios presentes en el corpus de tweets, y su correspondiente valoración, el paso definitivo es representar los datos y valoraciones obtenidas. Para ello, por cada temática se generan dos *rankings*, uno para valoraciones positivas y otro para negativas.

El score de valoraciones positivas se calcula de la siguiente manera:

$$\text{Score} = \#TotalApariciones * \frac{(\#ValoracionesPositivas - \#ValoracionesNegativas)}{max(\#ValoracionesPositivas, \#ValoracionesNegativas)}$$

Y el de valoraciones negativas es equivalente, solo que cambiando el orden de los operandos de la resta:

$$\text{Score} = \#TotalApariciones * \frac{(\#ValoracionesNegativas - \#ValoracionesPositivas)}{max(\#ValoracionesPositivas, \#ValoracionesNegativas)}$$

La función de ranking se basa en dar un peso de valoraciones positivas $\{\frac{(\#ValoracionesPositivas - \#ValoracionesNegativas)}{max(\#ValoracionesPositivas, \#ValoracionesNegativas)}\}$ o negativas $\{\frac{(\#ValoracionesNegativas - \#ValoracionesPositivas)}{max(\#ValoracionesPositivas, \#ValoracionesNegativas)}\}$ comprendido en el rango [-1, 1]. Por ejemplo, para el ranking de valoraciones positivas, si se tienen 10 valoraciones positivas y 0 negativas, en el numerador se obtiene 10-0 = 10, y en el denominador el máximo de 10 y 0, que sería 10. La división daría 1, que es el valor máximo. En el caso del ranking de valoraciones negativas, el orden de los operandos de la resta cambiaría, y por tanto el peso obtenido sería -1, cambiaría el signo. De este peso obtenido, se multiplica por la relevancia de la entidad, que aparece nombrada un mayor número de veces.

Con este ranking se consigue priorizar a las entidades o usuarios con mayor número de apariciones, y por tanto más relevantes en la temática, haciendo una distinción entre positivos y negativos.

A partir de los rankings construidos, la representación consiste en un documento HTML que muestra el porcentaje de valoraciones positivas, negativas y neutras para cada entidad o usuario, con enlace a los tweets en los que se encuentra la entidad o usuario para cada una de las valoraciones.

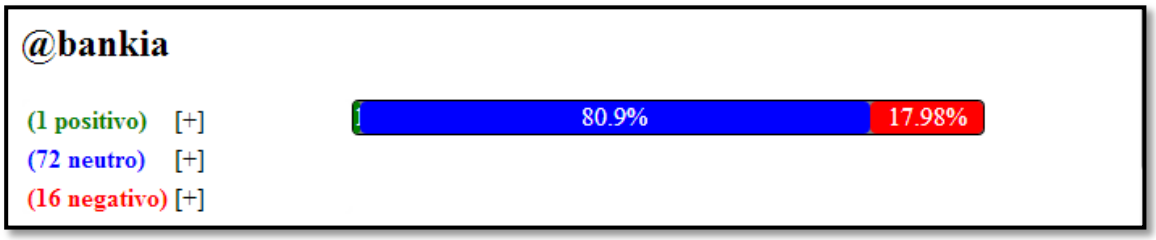


Figura 17. Porcentaje de valoraciones positivas, neutras y negativas para una entidad

Dado que hay números que pueden ser ilegibles, como el caso de la figura, se ha creado un *tooltip*, que al pasarlo por encima del porcentaje muestra el número completo, de manera que sea legible.

7.2 Representación de tweets

La representación de los tweets se basa en el formato HTML que proporciona Twitter para embeber los tweets en una página Web. Y que básicamente hace uso del link al tweet, basándose en el nombre del usuario que lo generó y el identificador del tweet. Además de referenciar al archivo javascript que da el aspecto del tweet y sus funcionalidades.

```
<blockquote class="twitter-tweet"><a
href="https://twitter.com/bolsamania/statuses/351987481186545664">2013-07-
02</a></blockquote> <script async src="http://platform.twitter.com/widgets.js"
charset="UTF-8"></script>

<b style="color: blue;">impacto</b>

<b style="color: red;">negativo</b>

<hr/>
```

Junto a los tweets se incorporan los lemas que se han tenido en cuenta para la valoración: Representándose los lemas calificados como negativos en color rojo, los positivos en verde, y los neutros en azul. Para el ejemplo del código HTML anterior, el resultado es el siguiente:

Figura 18. Tweet con lemas asociados a su valoración

7.3 Representación de entidades

En adición se han creado dos *tag clouds*. Un tag cloud es una nube de palabras, en este caso de entidades, en el que el tamaño de las palabras es mayor para las palabras que aparecen con mayor frecuencia, en este caso, cuantas más veces se mencionen en la temática.

El primer tag cloud contiene con el top 10 de usuarios Twitter mencionados, y el segundo el top 10 de entidades nombradas. El color que represente a la entidad será rojo, si ha sido extraído del ranking de valoraciones negativas o verde si ha sido extraído del ranking de valoraciones positivas. Para cada entidad de la nube se muestra un mensaje emergente o tooltip con las categorías extraídas de Wikipedia, y se enlaza, en el caso de los usuarios Twitter a la página principal de Twitter del usuario, y en el caso de la entidad nombrada, se enlaza a la búsqueda realizada en Google de la propia entidad.

Figura 19. Tooltip de categorías para la entidad BBVA

7.4 Salida final del sistema

Como resultado final para una temática dada se obtiene una página Web estática que puede ser accedida desde cualquier máquina, ya que el contenido es generado en un servidor de Apache. Dicha página Web está compuesta por los elementos mencionados anteriormente y organizada en cinco *frames* o marcos de la siguiente manera:

Figura 20. Salida final del sistema para la temática Bancos

En donde (1) corresponde al tag cloud de usuarios Twitter más mencionados en la temática y (2) al tag cloud de entidades nombradas el mayor número de veces. El frame (3) corresponde al ranking de valoraciones positivas, y el frame (4) al ranking de valoraciones negativas. En los frames de valoración, el icono '[+]' muestra en el frame (5), el de la derecha del todo los tweets relacionados con el usuario o entidad y valoración, con un máximo de cuarenta tweets.

8. Análisis de los datos obtenidos

En este capítulo se darán algunas de las salidas de cada una de las temáticas, mostrando el funcionamiento del sistema. Cada uno de los apartados corresponderá con las temáticas tratadas por el sistema. La representación, como se ha dicho, es de alguna de las salidas, ya que el mostrar la valoración para cada una de las entidades y cada una de las temáticas sería intratable.

A partir de aquí, se referirá a entidad, tanto a una entidad nombrada como a un usuario Twitter.

8.1 Reestructuración del sistema financiero en España

Figura 21. Salida del sistema para la temática Reestructuración del sistema financiero en España

En la temática Reestructuración del sistema financiero en España se observa una gran cantidad de valoraciones calificadas como neutras, no habiendo una tendencia clara a entidades positivas o negativas.

Como se ve en la figura 21, las entidades más relevantes, las de mayor número de veces mencionadas, tienen un alto porcentaje de calificaciones neutras. Fijándose en los números se ve mejor esta valoración, por ejemplo para la entidad Primaderiesgo, de 39 menciones en total, solamente una mención es positiva y el resto neutras, frente a esto, no se puede concluir que sea una entidad positiva.

La conclusión es de valoración neutra en la temática frente a los datos vistos, además, la mayoría de tweets generados son informativos.

8.2 Bancos

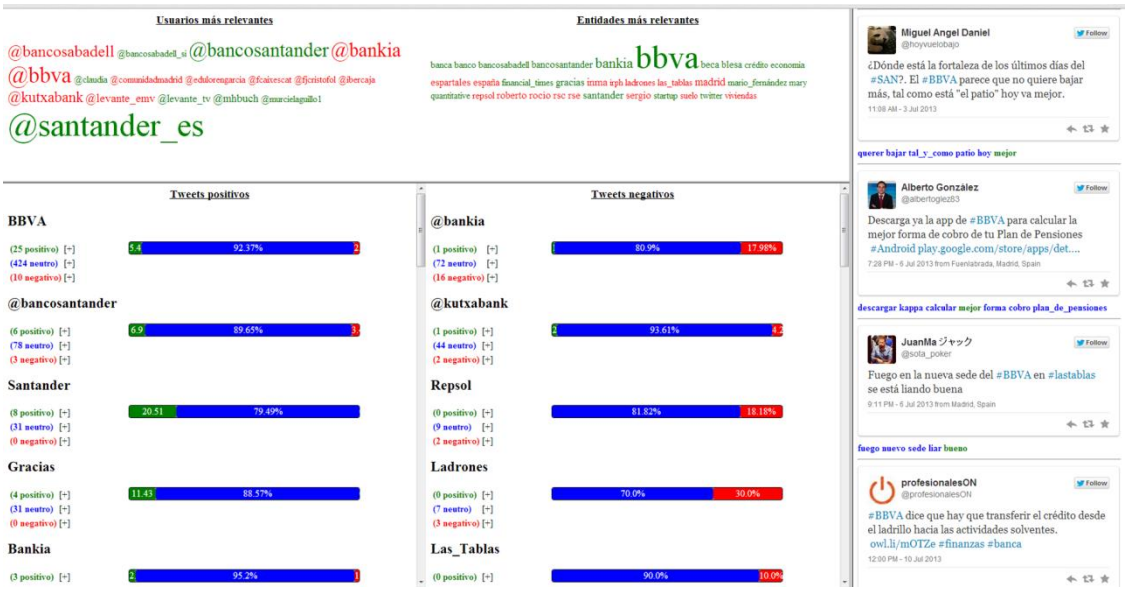


Figura 22. Salida del sistema para la temática Bancos con tweets positivos de la entidad BBVA

Como se puede ver en el ranking de la figura 22, la entidad mejor valorada a partir de los datos obtenidos es BBVA, dado que es una entidad bastante presente en la plataforma Twitter, con un total de 459 menciones, donde el número de valoraciones positivas, un 5.45%, supera al número de valoraciones negativas, un 2.18%, se puede correctamente como positiva debido a la gran cantidad de muestreo. A la derecha de la imagen anterior se pueden ver varios de los tweets que califican a la entidad como positiva.

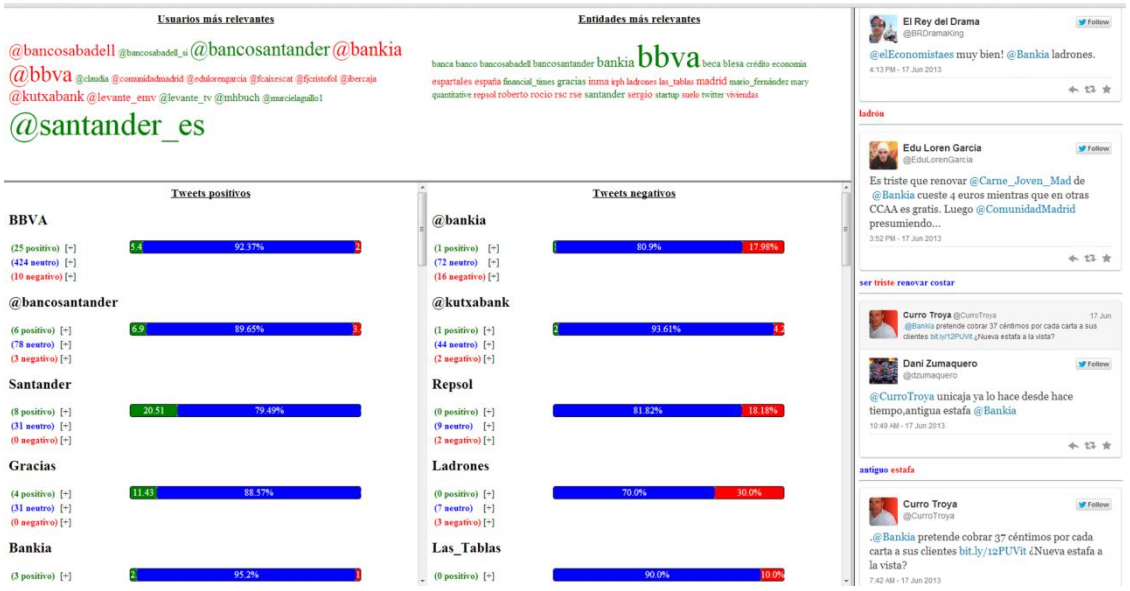


Figura 23. Salida del sistema para la temática Bancos con tweets negativos de la entidad @bankia

En cuanto a valoraciones negativas, se puede ver que @bankia y @kutxabank son de las peor valoradas. Aunque @bankia con mayor fortaleza de negatividad, un 17.98% de menciones negativas, frente a @kutxabank, un 4.26% de calificaciones negativas.

Fijándose en los rankings, destaca la incongruencia de la entidad Bankia tomada como positiva, y @bankia como negativo.

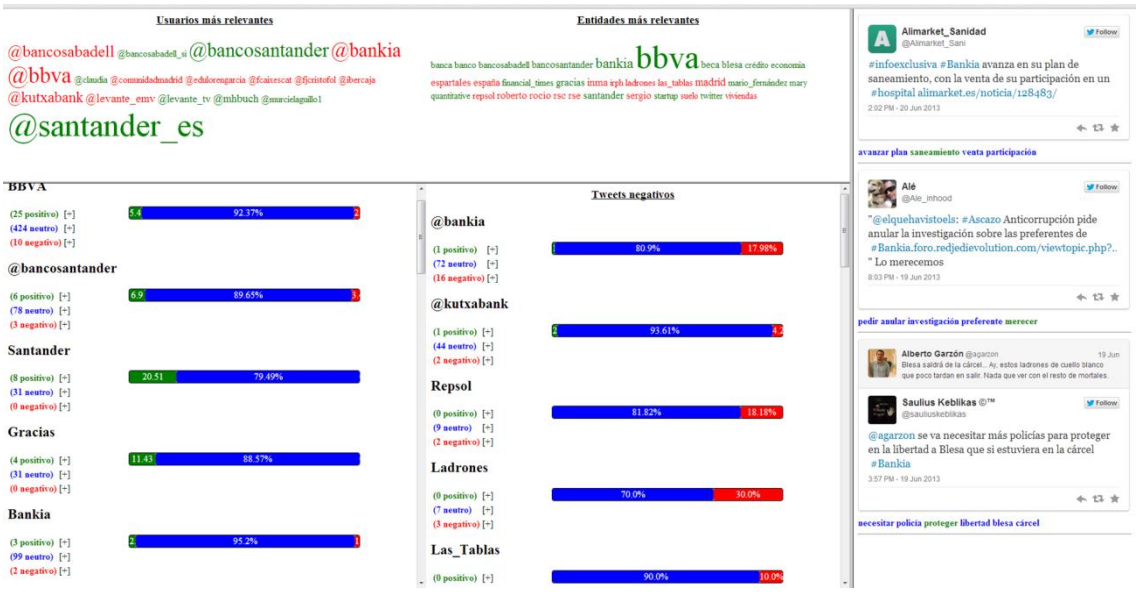


Figura 24. Salida del sistema para la temática Bancos con tweets positivos de la entidad Bankia

Viendo los tweets que califican a la entidad Bankia como positiva en la figura 24, se podría decir que es un falso positivo, dado que los pocos tweets que aparecen contiene ironía: "Lo merecemos", o son noticias informativas que poseen lemas tratados como positivos. De todas formas el porcentaje de valoración positiva es bastante bajo, 2.88%.

8.3 Desahucios

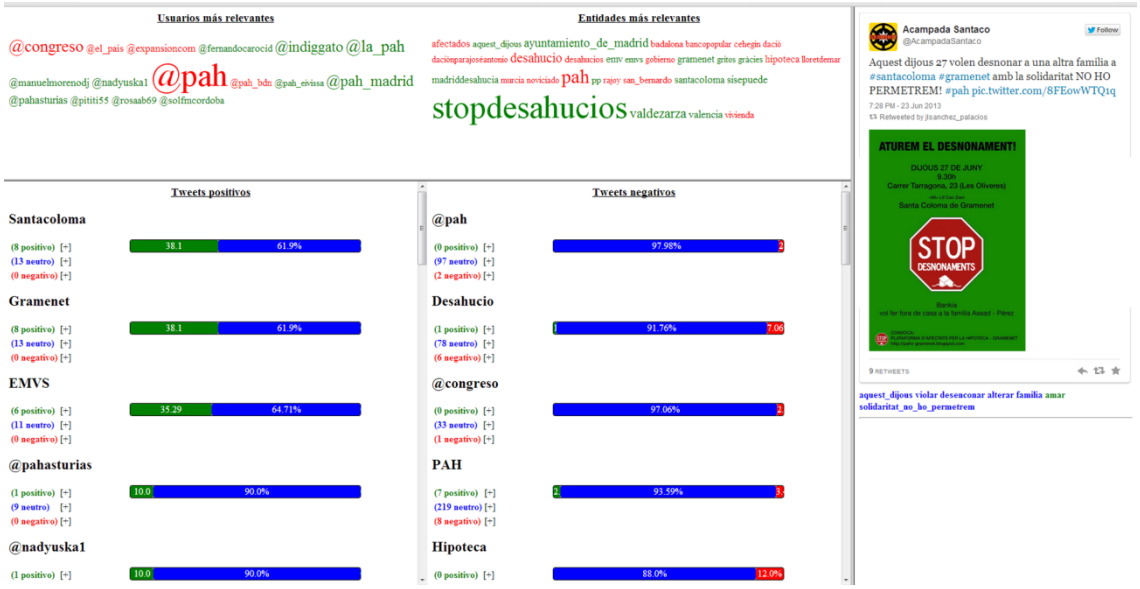


Figura 25. Salida del sistema para la temática Desahucios con tweets positivos de la entidad Santacoloma

En este caso destaca la problemática de la obtención de tweets en catalán, como se puede ver en la figura 25, lo que lleva al corrector ortográfico a interpretar palabras en catalán como "amb" a una corrección como "amar".

Figura 26. Salida del sistema para la temática Desahucios con tweets negativos de la entidad @pah

En el caso de la calificación de @pah ocurre lo mismo, se corrige "pot" como peor.

Dado que esta temática es la que menor número de tweets contiene, las muestras recogidas pueden no ser consistentes, y la gran cantidad de tweets en catalán hacen que se establezca el análisis como erróneo.

8.4 Preferentes

Figura 27. Salida del sistema para la temática Preferentes con tweets negativos de la entidad Bankia

En la temática de las preferentes, representado en la figura 27, se puede ver una gran cantidad de valoraciones neutras. En el caso del ranking de valoraciones positivas no se puede llegar a una fuerte conclusión de positivimismo, en su mayoría las valoraciones son neutras. En las entidades con un número de menciones superior a 10, Novagalicia y Liberbank, el número de menciones neutras es significativamente superior, en ambos casos solamente hay una mención como positiva. Observando los datos del ranking de valoraciones negativas, se puede concluir que la entidad pero valorada es Bankia , con un 6.96 % de valoraciones negativas, que se muestra junto a la entidad Afectados, pero viendo los tweets, se ve que las valoraciones de sentimiento negativo pertenecen a la entidad Bankia.

8.5 Educación

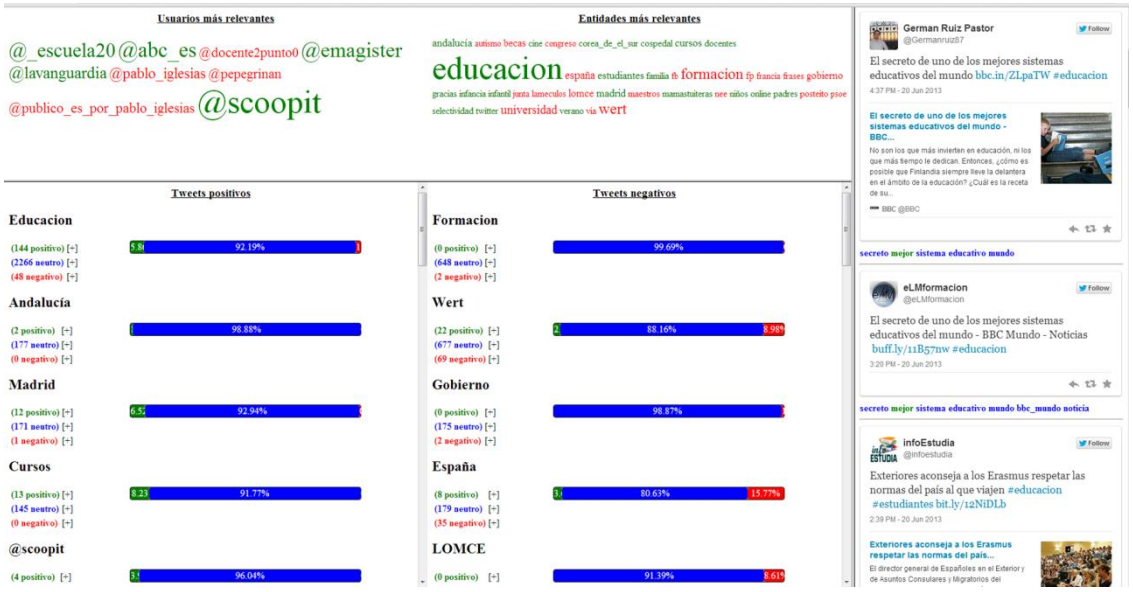


Figura 28. Salida del sistema para la temática Educación con tweets positivos de la entidad Educación

En la temática Educación, representada en la figura 28, se puede ver en el top del ranking de valoraciones positivas entidades como Educación y Cursos, así como Formación en el caso de valoraciones negativas. Lo que lleva a pensar que quizás los criterios de búsqueda no hayan sido los correctos por no ser demasiado específicos. Esto se confirma al ver los tweets de la entidad Educación, los tweets hablan de la educación, pero no tienen que ver concretamente con la educación en España.

Aquí es donde surge la importancia de una buena elección de los criterios de búsqueda. La valoración es correcta, pero no la obtención de tweets de la temática.

8.6 Sanidad

Figura 29. Salida del sistema para la temática Sanidad con tweets negativos de la entidad @igonzalezpp

En la temática Sanidad, representada en la figura 29, no se destacan entidades calificadas como positivas, en su mayoría, el ranking de valoraciones positivas el sentimiento vertido es neutro, en la mayor parte de los casos el porcentaje de valoraciones neutras supera el 95%. En cuanto al ranking de valoraciones negativas, se puede observar que la entidad peor valorada es @igonzalezpp con un mayor porcentaje de valoraciones negativas, un 15.28%.

9. Discusión

En este capítulo se dará una reflexión del trabajo realizado, así como las posibles mejoras y trabajo futuro. Concretamente, en el apartado 9.1 se expondrán las conclusiones personales sobre el trabajo fin de grado. En el apartado 9.2 se dará una visión de los problemas encontrados. Finalmente, en el apartado 9.3 se hará una propuesta de trabajos futuros en base al realizado, y que supondrían mejoras en la aplicación.

9.1 Conclusiones

En este trabajo de fin de grado se ha diseñado y desarrollado una herramienta capaz de identificar entidades nombradas en la plataforma Twitter, y analizar las valoraciones o sentimientos vertidos sobre las mismas, concluyendo con un resumen del análisis realizado a través de una interfaz Web.

Se ha tratado una temática novedosa y en auge, que es el análisis de sentimiento, donde no hay una solución definitiva. Por ello, en este proyecto se ha tratado de dar una propuesta de análisis de sentimiento, añadiendo como elemento diferenciador respecto a otras herramientas de sentimiento en Twitter el reconocimiento de sentimiento sobre entidades y usuarios.

El sistema no está cerrado, ya que continúan surgiendo nuevas ideas de mejora que potenciaría los resultados del análisis, que se aplicarían como trabajo futuro. Así mismo, se plantea una mejora de la interfaz gráfica, a la que no se ha dedicado el mismo esfuerzo que al análisis de opinión, el cual ha llevado la mayor parte de desarrollo del proyecto.

9.2 Problemas encontrados

Desde un punto de vista no técnico, las principales dificultades encontradas han sido el desconocimiento y novedad para el autor de la temática y tareas del proyecto, que tratan sobre el Procesamiento del Lenguaje Natural y la Minería de Opinión y el Análisis de Sentimientos. Éstas han requerido un esfuerzo de aprendizaje adicional, y una puesta en práctica de los nuevos conocimientos adquiridos.

En cuanto a problemas técnicos encontrados, se destaca la poca cantidad de software dedicado al Procesamiento del Lenguaje Natural para el idioma español. Aunque el software de PLN utilizado en este trabajo fin de grado, Freeling, es bastante potente y usable para el reconocimiento de entidades, con la complicación añadida de la escritura en Twitter, con lo que se ha necesitado un tratamiento previo del texto antes de ser analizado por la herramienta PLN.

Otro de los problemas encontrados es que en Twitter aunque se ajusten los criterios de búsqueda para tweets en español, se obtienen tweets con idiomas, lenguas o dialectos regionales, como el catalán, el gallego y el vasco. Este problema es debido a que usuarios registrados con idioma español escriben en su propia lengua regional, aunque en Twitter hayan registrado su lengua como el español. Esto provoca fallos en el análisis del texto.

Finalmente, cabe destacar la ironía como problema muy difícil a abordar, algo únicamente captable por el ser humano, ya que trata un tono "burlón", que una computadora no es capaz de comprender simplemente con el análisis de las palabras presentes en un texto.

9.3 Trabajo futuro

Como trabajo futuro se abre la posibilidad a varios proyectos. Uno de ellos puede ser un corrector mejorado al desarrollado en este proyecto, que contemple la escritura informal y el tratamiento de expresiones acotadas en la escritura, como: "xke", "k", "x?", etc. que son complejas de trasladar al lenguaje que podría llamarse formal. Lo que implicaría la creación de un gran diccionario, pudiéndose basar en proyectos ya creados de traducción, en los que usuarios voluntarios proponen la "traducción" de las expresiones. A todo esto, se le podría incorporar un corrector ortográfico basado en la posición de las teclas. Por ejemplo, si en una palabra se ha introducido una 'i', y se quería haber escrito una 'o'.

Otro posible proyecto sería el desarrollo de una herramienta capaz de detectar idiomas, que este caso de la aplicación desarrollada sería útil para la detección de idiomas regionales. Y que a su vez abre la opción a la creación de un traductor de idiomas multilingüe.

Como se puede ver, las opciones de desarrollo en el área del procesamiento del lenguaje son muy amplias, a la par de complejas.

Referencias

Apuntes de la asignatura "Extracción y Recuperación de Información", del Máster Universitario en Computación, de la Facultad de Informática de la Universidad de La Coruña.

http://www.grupolys.org/docencia/eri/

Libro on-line "*Introduction to Information Retrieval*", de Christopher D. Manning, Prabhakar Raghavan y Hinrich Schütze.

http://nlp.stanford.edu/IR-book/

Artículo académico "Minería de opiniones basada en características guiada por ontologías", de la Universidad de Alicante.

http://rua.ua.es/dspace/bitstream/ç10045/16947/1/PLN_46_11.pdf

Artículo de investigación "*TwiNER: Named Entity Recognition in Targeted Twitter Stream*", de IBM.

http://researcher.ibm.com/researcher/files/us-heq/sigir12twiner.pdf

Resumen del Proyecto fin de carrera "Análisis de sentimientos en Internet para decisiones estratégicas", de Carlos González Dulanto.

http://www.iit.upcomillas.es/pfc/resumenes/5049b5878939e.pdf

Tesis "Clasificación de Entidades Nombradas utilizando Información Global", de Carolina Rocío Sánchez Pérez.

http://ccc.inaoep.mx/~mmontesg/tesis%20estudiantes/TesisMaestria-CarolinaSanchez.pdf

Anexo A. Tecnologías empleadas

En este anexo se presentarán las tecnologías utilizadas para el desarrollo del sistema de extracción de entidades y análisis de contenidos.

Twitter API

La API de Twitter[16] es una interfaz que permite interactuar con los de datos de Twitter. Concretamente se basa en tres APIs:

- **Streaming API**[17]. La API de *streaming* que ofrece Twitter proporciona acceso de baja latencia a los datos generados en tiempo real. Esto requiere mantener abierta una conexión HTTP persistente con los servidores de Twitter.
- **REST API**[18]. El protocolo REST (*Representation State Transfer*) es un protocolo cliente/servidor sin estado donde cada mensaje HTTP contiene toda la información necesaria para entender la petición. En Twitter es utilizado para el acceso a los datos. Todas las operaciones que se pueden realizar vía Web, como obtener tweets generados por un usuario específico, seguidores de un usuario, etc., pueden ser accedidos a través de esta API.
- **Search API**[19]. La API de búsqueda ofrece una colección de tweets relevantes que coinciden con un criterio de búsqueda. Se ha de tener en cuenta que no todos los tweets generados en Twitter se han indexados en el sistema desarrollado, pues el tamaño del índice sería intratable para los recursos disponibles. Así, el índice usado por el sistema se reduce a tweets con una latencia máxima de entre 6 y 9 días.

Python-Twitter

Python-twitter[20] es una librería desarrollada en Python que permite fácil acceso a las funciones proporcionadas por la API de Twitter. Tiene acceso a las funcionalidades de las tres APIs proporcionadas por Twitter tratadas en el apartado anterior.

Una de las mayores ventajas de esta librería es la simplicidad en la autenticación con la API de Twitter, que se realiza en una única línea de código.

MySQL

MySQL[21] es un sistema de bases de datos de código abierto muy popularizado, con soporte para plataformas Windows y Linux. Se ejecuta sobre un servidor gestor de base de datos que provee acceso multiusuario a varias bases de datos.

[16] https://dev.twitter.com/
[17] https://dev.twitter.com/docs/streaming-apis
[18] https://dev.twitter.com/docs/api/1.1
[19] https://dev.twitter.com/docs/api/1.1/get/search/tweets
[20] https://github.com/bear/python-twitter
[21] http://www.mysql.com/

Este sistema no trae instalado por defecto ninguna interfaz GUI para la gestión o acceso a los contenidos de las bases de datos. Los usuarios pueden utilizar por defecto las herramientas de línea de comandos para el acceso y gestión. Sin embargo, existen herramientas que proveen de interfaz gráfica a la gestión y acceso a los datos, como Tora y SQLyog.

JDBC y MySQLdb

JDBC y MySQLdb son interfaces para la conexión a un servidor de bases de datos MySQL. En el caso de MySQLdb dicha interfaz está implementada en lenguaje de programación Python. Por ello ha sido utilizado en el crawler de Twitter, que también se ha desarrollado en Python. JDBC (*Java Database Connectivity*) está desarrollado en Java, y es utilizado por el sistema implementado para la gestión de las entidades reconocidas en los tweets, en los resultados del análisis de opinión y en la de otros datos recolectados.

Ambas interfaces proveen de acceso, lectura y escritura a la base de datos MySQL del sistema, únicamente diferenciándose del lenguaje de programación en el que están implementados.

FreeLing

FreeLing[22] es una librería *open source* para el procesamiento multilingüe que proporciona una amplia gama de funcionalidades de análisis de lenguaje natural para varios idiomas.

El proyecto FreeLing se inició desde el centro TALP[23] de la Universidad Politécnica de Cataluña para hacer más disponible el uso de recursos y herramientas básicos de Procesamiento del Lenguaje Natural, y así "*esta disponibilidad debería posibilitar avances más rápidos en proyectos de investigación y reducir costes en aplicaciones industriales de PLN*".

El proyecto está estructurado como una librería que puede ser invocada desde cualquier aplicación de usuario que requiera servicios PLN. Además, ofrece una amplia variedad de APIs para varios lenguajes de programación: Java, Perl, PHP, Python y Ruby.

Esta librería es una parte importante del sistema desarrollado, ya que proporciona los recursos de PLN necesarios.

EuroWordNet

EuroWordNet[24] es un recurso léxico integrado con Freeling, pero para hablar de EuroWordNet hay que referirse previamente a **WordNet**[25].

WordNet es una gran base de datos léxica de términos en inglés. Sustantivos, verbos, adjetivos y adverbios se agrupan en torno en conjuntos de sinónimos cognitivos (*synsets*), cada uno expresando un concepto distinto de un término.

[22] http://nlp.lsi.upc.edu/freeling/
[23] http://www.talp.upc.edu/
[24] http://www.illc.uva.nl/EuroWordNet/
[25] http://wordnet.princeton.edu/

EuroWordNet es un recurso multilingüe con varios idiomas europeos (holandés, italiano, español, alemán, francés, checo y estonio), que se basa en los recursos proporcionados por la base de datos de WordNet, que Freeling utiliza como diccionario para etiquetado gramatical, PoS.

SentiWordNet

SentiWordNet[26] es un recurso léxico utilizado en minería de opinión, distribuido como un archivo de texto. Está basado en el diccionario WordNet, y por cada entrada en el diccionario asigna puntuaciones de sentimiento positivas, negativas y neutras.

La implementación de este diccionario de polaridades se llevó a cabo mediante una compleja combinación de métodos de propagación y clasificadores. Por tanto, no es un recurso absoluto como WordNet, que fue compilado por seres humanos, pero ha demostrado ser bastante útil.

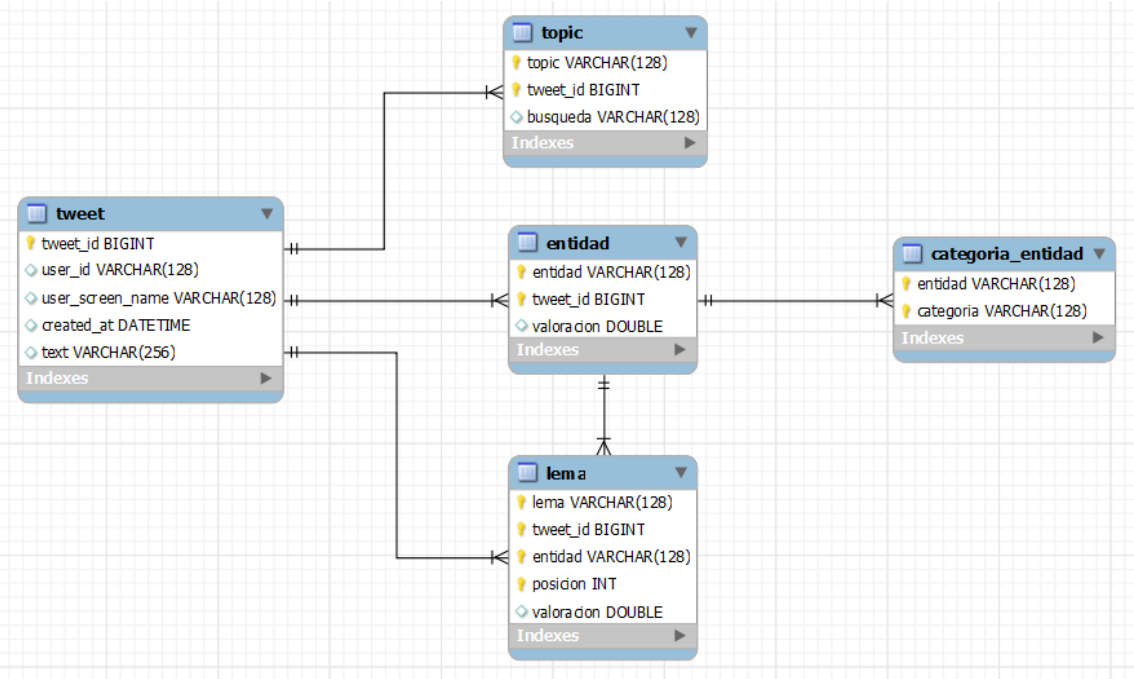
Aspell

GNU Aspell[27] es un corrector ortográfico libre y de código abierto. Utiliza diccionarios de distintos idiomas que han de estar instalados en el equipo, mediante lo cual realiza la comprobación de los términos del texto a corregir. Por cada una de las palabras que detecte como mal escritas, el programa propone una lista de palabras de sustitución.

[26] http://sentiwordnet.isti.cnr.it/
[27] http://aspell.net/

Anexo B. Diagrama entidad-relación de la base de datos

En este anexo se describen las tablas de la base de datos del sistema desarrollado, cuyo diagrama de entidad-relación se muestra en la siguiente figura.



tweet	
tweet_id	Identificador único de un tweet.
user_id	Identificador único del usuario autor del tweet.
user_screen_name	Nombre de la cuenta del usuario autor del tweet.
created_at	Timestamp del tweet.
text	Texto del tweet.

topic	
topic	Temática del tweet descargado
tweet_id	Identificador único del tweet
busqueda	Términos de búsqueda para la descarga

entidad	
entidad	Nombre de la entidad
tweet_id	Identificador del tweet al que pertenece

categoría_entidad	
entidad	Nombre de la entidad
categoría	Categoría de la entidad

lema	
lema	Lema obtenido de un tweet
tweet_id	Identificador único del tweet
entidad	Nombre de la entidad a la que pertenece el lema
posicion	Posición del texto en el que aparece el lema
valoracion	Score de valoración de sentimiento del lema